李英杰　著

数据挖掘算法及在视频分析中的应用

SHUJU WAJUE SUANFA JI ZAI SHIPIN FENXI ZHONG DE YINGYONG

中国水利水电出版社
www.waterpub.com.cn

数据挖掘算法及在视频分析中的应用

李英杰 著

中国水利水电出版社
www.waterpub.com.cn

内 容 提 要

随着网络与计算机的发展，可利用的数据量日益增大，数据的形式更多样化，这对数据挖掘算法的研究和数据挖掘与领域知识、技术的融合都提出了新的挑战。本书在分析数据挖掘相关概念和相关技术研究现状基础上，阐述了围绕数据挖掘中的分类、特异数据挖掘、关联规则等任务中经典算法的改进研究。继而阐述了数据挖掘算法在计算机视觉领域应用的研究工作。每部分研究工作均详细描写了背景、问题、研究思路、实验结果、结论与总结等。各部分工作相关，又独成体系，可读性好。

本书可作为高等学校数据分析类课程的补充资料，也可供相关方向的研究生及专业科技工作者参考。

图书在版编目（CIP）数据

数据挖掘算法及在视频分析中的应用 / 李英杰著
. -- 北京 : 中国水利水电出版社, 2014.5
ISBN 978-7-5170-1997-8

Ⅰ. ①数… Ⅱ. ①李… Ⅲ. ①数据采集 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2014)第096349号

书　　名	**数据挖掘算法及在视频分析中的应用**
作　　者	李英杰　著
出版发行	中国水利水电出版社 （北京市海淀区玉渊潭南路1号D座　100038） 网址：www.waterpub.com.cn E-mail：sales@waterpub.com.cn 电话：（010）68367658（发行部）
经　　售	北京科水图书销售中心（零售） 电话：（010）88383994、63202643、68545874 全国各地新华书店和相关出版物销售网点
排　　版	中国水利水电出版社微机排版中心
印　　刷	三河市鑫金马印装有限公司
规　　格	170mm×240mm　16开本　8印张　152千字
版　　次	2014年5月第1版　2014年5月第1次印刷
印　　数	0001—2000册
定　　价	**18.00**元

前言

数据挖掘是高级数据分析工具，其任务包括频繁项集挖掘、关联规则挖掘、聚类、分类、特异数据挖掘、时间序列挖掘等。随着信息技术的发展，人类进入了大数据时代，数据分析的需求日益增长，并且多样化，这对数据挖掘算法的研究和数据挖掘与领域知识、技术的融合都提出了新的挑战。

本书在分析数据挖掘相关概念和相关技术研究现状基础上，阐述了围绕数据挖掘算法和其在计算机视觉领域应用的研究工作。研究工作包括一种基于聚类的全局特异数据挖掘算法、利用必要关联规则提高分析精度的方法、一种基于外观的视频中行为识别特征的提取方法，以及基于差分序列的外观特征和光流特征对行为识别效率的评估。书中详细阐述了每项工作的背景、问题、解决思路、算法、实验结果、结论等。

书中内容是作者阶段性研究工作的汇总与总结。文中工作得到浙江农林大学人才启动经费支持（2010FR070），部分工作是作者博士期间完成的，感谢导师尹怡欣教授的指导，感谢博士同学们的帮助。本书撰写过程中也得到了一些专家和学者的指导，还借鉴和引用了大量国内外有关教材、专著、论文、标准等资料，在此，谨向他们表示衷心的感谢！

研究工作是无止境的，书中含有一些个人观点，加之作者水平有限，书中的缺点和错误在所难免，恳请各界读者提出宝贵意见。

著　者

2014 年 3 月

目　录

第1章 绪 论

计算机技术与网络技术的快速发展和软件技术的广泛应用使数据大量积累，而计算机硬件技术的不断进步又使大规模数据的集中存储成为可能。剧增的数据中有可能隐藏着许多重要的信息，人们希望能够对拥有的信息进行更高层次的分析，产生了对数据分析工具的强烈需求。数据挖掘是高级数据分析工具，它可以帮助人们从数据库特别是数据仓库的相关数据中提取出所感兴趣的知识、规律或更高层次的信息，而且也可以帮助人们从不同程度上去分析它们；它不仅可以用于描述过去数据的发展过程，而且还能进一步预测未来的发展趋势。

数据挖掘成为热点研究领域已有相当长的时间，取得了阶段性的成果，目前国内外都有一些集成了数据挖掘方法的数据分析工具。如 SPSS 公司的 Clementine SPSS for Windows，其中集成了神经网络、回归、因子分析、决策树、聚集、关联规则、规则归纳、单调回归和 OLAP 环境等技术[1]；SAS 公司的 Enterprise Miner 集成了数据挖掘的多种工具[1]；一些关系数据库产品中都有数据挖掘和联机数据分析工具；国内也有已经获得软件著作权的产品，如 DMINER[2]；另外还有许多专门开发数据挖掘项目的公司。但是，这绝不意味着数据挖掘的研究已经结束，相反它留给研究者丰富的理论和实践课题[3]。

数据挖掘继承和发展了相关基础学科的理论与技术，如基础统计学、概率论与数理统计、机器学习等，探索出了独具特色的理论体系。但是，随着数据挖掘本身技术和相关技术的发展、随着应用日益广泛及新需求的推动，新的挖掘理论研究是必需的。新理论会促进新算法的产生，同时，现有算法的效率、适用性和实用性也有待提高。

数据挖掘一般有数据准备、知识挖掘和模式评估 3 个阶段。数据挖掘系统基本构架已趋于明朗，但不同领域、不同数据类型、不同知识表达模式下的应用，其具体的实现机制、技术路线等方面仍有待于结合实际进行深入研究。由于用户事先不知道数据源中潜在的知识，在数据挖掘的各个阶段与用户的交互都是必需的，良好的交互方式及清晰友好的可视界面是系统成功应用的前提，所以交互的时机、适应交互式的有效算法、交互过程的可视化及与其适应的构架研究都是数据挖掘领域的重要课题。

互联网的发展加速了信息的爆炸式增长。如何有效地获取有用的互联网信息与知识是数据挖掘的重要目标。同时，互联网为数据挖掘提供了良好的挖掘环境与挖掘对象，且其挖掘结果可获得可见的回报，Web 挖掘、文本挖掘成为目前研究的热点[4]。在日本成立了世界性科学组织 Web 智能协会（Web Intelligence Consortium，WIC），并且在 2003 年成立了 WIC 波兰中心及 WIC 英国中心[6]，致力于网上智能的研究与推广。Web 挖掘研究是网络智能的基础[8]。Web 挖掘被定义为：从与 WWW 相关的资源和行为中抽取感兴趣的、有用的模式和隐含信息[4]。Web 挖掘一般分为 3 类：Web 内容挖掘、Web 结构挖掘和 Web 使用挖掘。Web 挖掘可以在获取和分析网上信息过程中的很多方面发挥作用，如：确定权威页面、页面分类、页面信息抽取、挖掘用户的访问方式、智能查询和网上情报探测等。

搜索引擎是大多数人搜寻网上信息的首选。搜索引擎主要是以关键字匹配的方式返回可能是用户要寻找的网页的链接。搜索引擎一般返回大量的链接，往往让用户无所适从，对网上信息进一步分析成为迫切需求。受此驱动，很多学者针对网络智能的各方面进行了研究。Malik 等详细研究了挖掘 Web 上的特异内容、特异结构、特异使用的任务、方法和构架[11]。Lee R. S. T. 等开发了一个多智体构架的网上购物信息挖掘系统[17]。构架集成了人脸识别、用户需求定义、模糊需求模式、模糊多智体协商、产品评估模式等多项技术，系统能较好地理解用户的需求，达到较高的匹配率。为了使计算机能有效地访问网络资源，很多学者致力于研究和开发语义 Web，语义 Web 环境下的信息检索和分析也在积极的研究中。但目前其研究和实验只限定在实验室的虚拟环境中[18]。为了更好地服务用户，搜索引擎不断地利用数据挖掘、模糊匹配等分析方法挖掘用户的意图、计算网页与用户提问的相似程度。阿里巴巴与雅虎在 2006 年 7 月推出了 Imatch 模糊匹配智能搜索，它可以根据用户搜索习惯和意图智能匹配相关搜索结果，贴近用户的实际需求[6]，类似的还有百度知道等[24]。另外，一些专业搜索引擎也纷纷推出，以提高搜索精度。如购物搜索、专业 DJ 搜索、IT 专业搜索、教育搜索等。目前的搜索智能主要来源于对用户已经提出过问题的挖掘，用户希望有能与之交流类似于专家互动过程的更智能的搜索引擎出现。情报是企业第四核心竞争力，网络已经成为各类情报的首要获取途径，人们迫切需要自动的系统来从网上持续、系统地搜集及分析信息，形成有价值的情报，以支持企业、商家及个人决策[22]。

数据挖掘也已经成为数据库理论和应用的一个重要方向。关系数据库标准查询语言 SQL 允许用户提出特定的数据检索要求，有力推动了关系数据库的应用和发展。数据挖掘的数据主要来源于数据库尤其是关系数据库，开发数据挖掘标准查询语言并与 SQL 无缝衔接是数据挖掘技术发展的必然要求。充分

利用SQL语言和关系数据库运行机制，提高挖掘算法的效率成为重要的研究方向。

大规模数字采集与存储设备的普及促进了空间数据与多媒体数据挖掘的研究，同时促进了数据分析技术与声音与图像处理技术的融合。智能设备、智能软件产品逐步走入了我们的日常生活。由于声音、图像处理技术的复杂性与现实应用的不确定性和复杂性，产品的智能还有待深入发掘。

另外，对复杂数据类型数据源的挖掘、增量式算法开发、不同技术在同一系统中的集成、对挖掘结果的评价、对数据私有权的保护与信息安全等均是有待深入研究的方向。

在学习和总结前人工作的基础上，深入研究数据挖掘算法在挖掘大型数据源时的效率问题，针对具体的应用对这些算法进行改进和创新，提出更有效的算法，并研究数据挖掘技术在智能视频监控中的应用，发现视频中运动物体、人的行为判断等的有效模式是本书的研究目标。通过前一阶段的研究，作者对这个领域的知识、方法、问题和解决思路已有较深入的理解，取得了阶段性成绩，同时，培养了作者思考、分析和解决问题及独立科研的能力。后续的研究，将主要集中在数据挖掘理论与算法在视频智能监控系统中的应用研究，争取在领域的某些点上有所突破，为数据挖掘技术的发展贡献一份力量。

第 2 章　数据挖掘基本概念与相关技术研究现状

数据挖掘是一个多学科交叉研究领域。它融合了数据库技术、人工智能、机器学习、统计学、知识工程、面向对象方法、信息检索、高性能计算以及数据可视化等最新技术的研究成果。数据挖掘正以一种全新的概念改变着人们利用数据的方式。20 世纪以来，数据库技术取得了决定性的成果并且已经得到广泛的应用。在关系数据库和事务数据库的基础上，已经对数据挖掘方法在许多领域进行了较为深入广泛的研究，包括数据流信息处理、空间信息处理、多媒体信息、时空信息以及移动对象信息等。

本书关注聚类、分类与特异数据挖掘的算法及它们的融合方法，同时关注 Web 挖掘和视频图像挖掘等应用方向。相关的概念、方法及应用现状将在本章阐述。

2.1　数据挖掘的基本概念

数据挖掘就是综合应用一系列先进的技术从大量数据中提取人们感兴趣的信息和知识，它们是隐含的、事先未知且潜在有用的概念、规则、规律及模式等。这个概念诠释了数据挖掘的 3 个要点：

(1) 数据挖掘要处理的数据量是巨大的。因此，高效率常常是数据挖掘算法研究的目标。

(2) 要挖掘的概念、规则、规律和模式是事先未知的，挖掘结果是否有效与领域知识、人们的兴趣和当时的背景有关。因此，建立客观通用的效果评价标准有难度。

(3) 理论上，从领域知识中无法总结出规律，而欲从大量数据中找到规律，以发现问题、指导后续工作的应用均可使用数据挖掘技术来解决。而在实际问题中，数据的获取方式、数据格式、数据噪声、模式的描述与解释多种多样，数据挖掘技术与领域知识、领域技术的融合、提升，是创新也是挑战。

数据挖掘的任务包括频繁项集挖掘、关联规则挖掘、聚类、分类、特异数据挖掘和时间序列挖掘等。数据挖掘的过程一般包括以下可能重复的过程[25]：

1）数据清洗：删除噪声数据和不一致数据。

2）数据集成：将多个数据源进行合并整理。

3）数据挑选：从数据库中选出与分析任务相关的数据。

4）数据转换：数据要被转换和整理，使其符合挖掘程序的格式。

5）数据挖掘：执行挖掘过程，获取数据模式。

6）模式评估：评估挖掘出模式的有效性。

7）知识表示：利用可视化和知识表示技术将挖掘结果呈现给用户。

数据挖掘的对象数据形式可能是关系数据库、数据仓库、事务数据库、对象关系数据库、多媒体数据库、时间空间数据库、文本文件、数据流及 Web 等[25]。典型的数据挖掘系统构架如图 2.1 所示[25]。

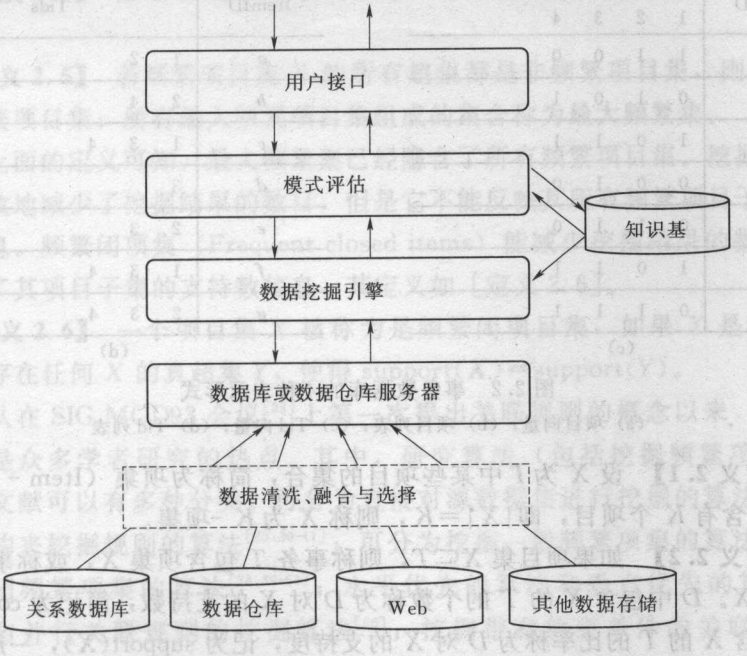


图 2.1 典型的数据挖掘系统构架

2.2 频繁项集和关联规则挖掘

2.2.1 问题描述

设 $I=\{i_1,i_2,\cdots,i_m\}$ 是 m 个不同项目的集合。D 是所有事务的集合（即事务数据库），每个事务 T 是一些项目的集合，T 包含在 I 中，即 $T\subset I$，并且每个事务可以用唯一的标识符 Tid 来标识。通常 D 是二维表，一维表示项

目，另一维表示每个 Tid 事务所具有的项目情况，它有 4 种表示形式：项目向量（Item－vector）、项目列表（Item－list）、Tid 向量（Tid－vector）和 Tid 列表（Tid－list），示例如图 2.2 所示[3,26,27]。

Tid	ItemID						
	a	*b*	*c*	*d*	*e*	*f*	*g*
1	1	0	1	0	0	1	0
2	1	1	0	0	1	0	1
3	0	0	1	1	1	1	1
4	0	1	1	0	0	1	1

(a)

Tid	ItemIDs
1	*a* *c* *f*
2	*a* *b* *e* *g*
3	*c* *d* *e* *f* *g*
4	*b* *c* *f* *g*

(b)

ItemID	Tid			
	1	2	3	4
a	1	1	0	0
b	0	1	0	1
c	1	0	1	1
d	0	0	1	0
e	0	1	1	0
f	1	0	1	1
g	0	1	1	1

(c)

ItemID	Tids
a	1 2
b	2 4
c	1 3 4
d	3
e	2 3
f	1 3 4
g	2 3 4

(d)

图 2.2　事务数据库的 4 种表示形式

(a) 项目向量；(b) 项目列表；(c) Tid 向量；(d) Tid 列表

【定义 2.1】 设 X 为 I 中某些项目的集合，简称为项集（Item－set）。如果 X 中含有 K 个项目，即 $|X|=K$，则称 X 为 K -项集。

【定义 2.2】 如果项目集 $X \subseteq T$，则称事务 T 包含项集 X，或称事务 T 支持项集 X。D 中包含 X 的 T 的个数称为 D 对 X 的支持数，简记为 count(X)。D 中包含 X 的 T 的比率称为 D 对 X 的支持度，记为 support(X)，一般用百分比表示。如果用 $|D|$ 表示 D 中 T 的总数，显然有：$\text{support}(X)=\dfrac{\text{count}(X)}{|D|}$。

【定义 2.3】 关联规则是形如：$X \Rightarrow Y$ 的蕴涵式，这里 $X \subset I$，$Y \subset I$，并且 $X \cap Y = \Phi$。称 X 为前件（antecedent），Y 为后件（consequent）。称 support（$X \Rightarrow Y$）＝support（$X \cup Y$）为此关联规则的支持度。关联规则 $X \Rightarrow Y$ 的置信度是包含 $X \cup Y$ 的事务数与包含 X 的事务数的比值，记为 confidence（$X \Rightarrow Y$），所以有：

$$\text{confidence}(X \Rightarrow Y)=\frac{\text{count}(X \cup Y)}{\text{count}(X)}=\frac{\text{support}(X \cup Y)}{\text{support}(X)} \tag{2.1}$$

给定最小支持度阈值 minsup 和最小置信度阈值 minconf，关联规则挖掘即在 D 中找出所有支持度≥minsup 和置信度≥minconf 的关联规则。

【定义 2.4】 若项集 X 的支持度不小于最小支持度，则称 X 为频繁项目集。若某一项目 i_m 满足最小支持度要求，则称 i_m 为频繁项目，所有频繁项目的集合称为频繁 1-项集，记为 L_1；满足最小支持度要求的 k-项集称为频繁 k-项集，所有频繁 k-项集的集合记为 L_k。

给定最小支持度阈值 minsup，频繁项集挖掘即在 D 中找出所有支持度≥minsup 的频繁项集。频繁项集具有如下两个性质：

【性质 2.1】 如果 X 是频繁项目集，那么 X 的任何非空子集都是频繁项目集。

【性质 2.2】 如果 X 是非频繁项目集，那么 X 的任何超集都是非频繁项目集。

【定义 2.5】 若频繁项目集 X 的所有超集都是非频繁项目集，则称 X 为最大频繁项目集；所有最大频繁项目集组成的集合称为最大频繁集。

由上面的定义可知，最大频繁集已经隐含了所有频繁项目集。挖掘最大频繁集有效地减少了挖掘结果的数目，但是它不能反映其所有频繁项目子集的支持数信息。频繁闭项集（Frequent closed items）能减少挖掘结果的数目，同时保留了其项目子集的支持数信息，其定义如［定义 2.6］。

【定义 2.6】 一个项目集 X 被称为是频繁闭项目集，如果 X 是频繁的，并且不存在任何 X 的真超集 Y，使得 support(X)＝support(Y)。

自从在 SIG MOD93 会议[28]上第一次提出关联规则的概念以来，关联规则一直是众多学者研究的热点。其中，研究算法（包括挖掘频繁项集的算法）的文献可以有多种分类。可分为直接对源数据集进行挖掘的算法和借助数据结构来挖掘规则的算法[29,30-41]；可分为挖掘一般频繁项集的算法和挖掘最大、闭频繁项集的算法[38,39-53]；水平优先的算法和垂直优先的算法[54]；还有分布并行关联规则的挖掘算法[56]；挖掘带有约束条件的关联规则算法[40,55-70]；关联规则增量更新挖掘算法[35]等。还有相当多的学者研究了关联规则挖掘与关系数据库紧密结合的问题[71,72-88]，关联规则的结果评价标准问题[81]，挖掘系统的构架、交互方式及可视化问题[83]等。但是，数据挖掘乃至关联规则挖掘技术还没有成熟，要达到实用推广还需要做长期艰苦的探索。

2.2.2 关联规则经典挖掘算法 Apriori

Apriori 算法[3]是单维、单层、布尔关联规则挖掘算法，是最简单形式的关联规则挖掘。该算法是挖掘产生布尔关联规则频繁项目集的经典算

法，对关联规则挖掘研究有着重要影响。首先，关联规则的挖掘被分为两个步骤：

(1) 找出满足 minsup 的所有频繁项集。

(2) 从频繁项集生成关联规则。

步骤 (1) 中，算法利用一个逐层搜索的迭代方法来完成频繁项目集的挖掘。具体的做法如下：首先访问一次数据库，找出频繁 1 -项集，记为 L_1；利用 $L_1 \times L_1$ 生成频繁 2 -项集候选集 C_2；访问一次数据库筛掉 C_2 中的非频繁项集，形成频繁 2 -项集 L_2；利用 $L_2 \times L_2$ 生成 3 -项集，根据 [性质 2.1] 和 [性质 2.2] 去掉 3 -项集中不可能频繁的项集实现剪枝，留下的作为 C_3；访问一次数据库筛掉 C_3 中的非频繁项集，形成频繁 3 -项集 L_3；如此不断地循环下去直至 C_k 或 L_k 空为止。

Apriori 算法中将由 $L_k \times L_k$ 生成 $k+1$ -项目集，并剪枝形成 C_{k+1} 的过程分离出来称 Apriori - gen 算法，如图 2.3 所示。

输入：L_k　// 频繁 k -项集

输出：C_{k+1}　// 候选 $k+1$ -项集

Apriori - gen 算法：

$C_{k+1} = \Phi$;

For each I$\in L_k$ do

for each $J \in L_k$ and $J \neq I$ do

if I 与 J 中有 $k-1$ 项是相同的　then

$C_{k+1} = C_{k+1} \cup \{I \cup J\}$;

对 C_{k+1} 中所有项目集检查它的所有 k 子集是否全在 L_k 中，如果此项目集的某个 k 子集不在 L_k 中，则从 C_{k+1} 中去掉它；

Return C_{k+1};

图 2.3　Apriori - gen 算法

Apriori 算法调用 Apriori - gen，生成所有频繁项集，如图 2.4 所示。

Apriori 算法假定数据库驻留在内存中。数据库扫描的最大趟数等于最大的频繁项目集的基数加 1。生成了频繁项集 L 后，关联规则的生成变得非常直接。其算法命名为 ARGen，如图 2.5 所示。

许多情况下，Apriori 算法的产生——剪枝方法大幅度地压缩了候选项集的大小，具有较好的性能。但在数据库规模巨大、项目稠密的情况下，频繁 1 -项集很大和最终产生的频繁模式很长，Apriori 算法就需要大量的剪枝运算和多次扫描数据库，算法的效率会大大降低。Apriori 算法之后，学者们不断研究其改进算法及其他思想的关联规则挖掘算法，取得了很多成果。

输入：I　//项目集合
　　D　//事务数据库
　　s　//最小支持度
输出：L　//频繁项集
Apriori 算法：
　　$K=0$；
　　$L=\Phi$；
　　$C_1=\mathrm{I}$；
　　Repeat
　　　$K=K+1$；
　　　$L_k=\Phi$；
　　　For each $I_i \in C_k$ do
　　　　$\mathrm{count}_i=0$；　　//为每个 I_i 计数赋初值 0
　　　for each $t_j \in D$ do
　　　　for each　$I_i \in t_j$ do
　　　　　$\mathrm{count}_i=\mathrm{count}_i+1$；
　　　For each $I_i \in C_k$ do
　　　　If $\mathrm{count}_i \geqslant (s \times |D|)$ then
　　　　　$L_k=L_k \cup I_i$；
　　　$L=L \cup L_k$；
　　　$C_{k+1}=\mathrm{Apriori-gen}(L_k)$
　　Until $C_{k+1}=\Phi$；
　　Return L；

图 2.4　Apriori 算法

输入：D　//事务数据库
　　I　//项目集合
　　L　//频繁项目集
　　s　//最小支持度
　　a　//最小置信度
输出：R　//满足 s 和 a 的关联规则集合
ARGen 算法：
　　$R=\Phi$；
　　For each $\Gamma \in L$ do
　　　for each $X \subset \Gamma$　　and $X \neq \Phi$do
　　　　if $\dfrac{\mathrm{support}(\Gamma)}{\mathrm{support}(X)} \geqslant a$ then
　　　　　$R=R \cup \{X \Rightarrow \Gamma - X\}$；
　　Return R；

图 2.5　找出频繁项集 L 后生成关联规则算法

2.2.3　FP－growth 算法

在借助数据结构存放中间结果来挖掘关联规则的算法中，FP－growth 算法[89]效率较高，其中使用的树结构 FP－Tree 成为很多研究的基础。算法的重点还是在挖掘频繁项集的过程，分为两部分：

(1) 通过两次扫描事务数据库，把每个事务所包含的频繁项目按其支持度降序压缩存储到 FP－Tree 中。

(2) 通过递归调用 FP－growth 的方法来直接产生频繁项目集。这个过程不需要再扫描事务数据库，而仅在 FP－Tree 中进行查找即可，在整个发现过程中也不需产生候选项目集。

FP－Tree 由两部分构成：头表和树。头表中节点由 itemname、count 和 node _ link 3 个域组成，分别标识项目名、支持数和指向树中第一个同名节点的指针。树中节点由 itemname、count、parent _ link、node _ link 4 个域组成，分别标识项目名、计数、父节点的指针和指向树中下一同名节点的指针。通过以下例子来说明 FP－Tree 的构造过程。

【例 2.1】 (见文献 [89] 中的 [例 3.1])：设有如表 2.1 所示的前两列事务数据库。令最小支持数为 3。则 FP－Tree 的构造过程如下：

表 2.1　　**示例事务数据库**

Tid	Items	Frequent items (ordered)
100	f, a, c, d, g, I, m, p	f, c, a, m, p
200	a, b, c, f, l, m, o	f, c, a, b, m
300	b, f, h, j, o	f, b
400	b, c, k, s, p	c, b, p
500	a, f, c, e, l, p, m, n	f, c, a, m, p

(1) 访问一次数据库，找出所有频繁 1－项集，并按支持数降序排列，结果为：$<(f: 4), (c: 4), (a: 3), (b: 3), (m: 3), (p: 3)>$ (“:” 后面表示项目的支持数)。生成头表中的项，并将每项后的指针初始化为 “null”。为方便后面叙述，表 2.1 中在第 3 列列出了每个事务中按降序排列的频繁项 (Frequent items)。

(2) 创建树的根节点并标记为 null。第二次扫描数据库并构建 FP－Tree。

取到第 1 个事务；析出其中的频繁项；降序排列；在根下创建树的第一个分支：$<(f: 1), (c: 1), (a: 1), (m: 1), (p: 1)>$。

取到第 2 个事务；析出其中的频繁项；降序排列$<f,c,a,b,m>$。因为其前 3 项$<f,c,a>$与现在树中第一个分支的前 3 个节点重复，只需在树中把

这3个节点的计数分别加1变成<(*f*：2)，(*c*：2)，(*a*：2)>；再给节点<(*a*：2)>增加一个子分支<(*b*：1)，(*m*：1)>。

取到第3个事务；析出其中的频繁项；降序排列<*f*,*b*>。因为其第一项<*f*>与现在树中第一个分支的前一个<(*f*：2)>节点重复，只需树中把这3个节点的计数分别加1变成<(*f*：3)>；再给节点<(*f*：3)>增加一个子分支<(*b*：1)>。

取到第4个事务；析出其中的频繁项；降序排列<*c*,*b*,*p*>。因为其与现在树中的分支没有前端重复，所以给根节点增加一个子分支<(*c*：1)，(*b*：1)，(*p*：1)>。

取到第5个事务；析出其中的频繁项；降序排列<*f*,*c*,*a*,*m*,*p*>。因为其与现在树中最左分支完全重复，所以只需给这个分支的每个节点的计数加1，变为：<(*f*：4)，(*c*：3)，(*a*：3)，(*m*：2)，(*p*：2)>。

这个过程中随时将各新生成节点与头表中相应的项进行链接。最后生成的FP-Tree如图2.6所示。

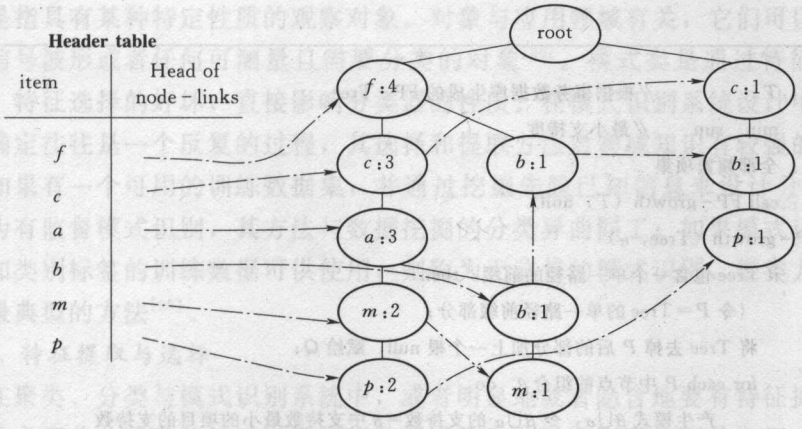


图2.6 [例2.1]生成的FP-Tree

FP-Tree的构造算法如图2.7所示。

可以看出，通过两次访问数据库，所有频繁项目的信息被存放到FP-Tree中。FP-growth算法以FP-Tree头表的倒序为计算顺序，从下至上递归搜索FP-Tree，逐步生成所有的频繁项集。其算法如图2.8所示。

FP-Tree在不损失事务信息的情况下，有效地压缩了数据库中数据，是迄今为止提出的较有效的数据结构，目前有许多研究都是基于这个结构的。

输入：TDB　　//事务数据库

　　min_sup　　//最小支持度

输出：T　　//生成的 FP-Tree

FP-tree 构造算法：

(1) 扫描一次数据库，生成频繁项目和其对应的支持数的集合 F，将 F 降序排列记为 Flist。

(2) 生成根节点 root，并标记为“null”，这时 T 中只含一个根节点。对数据库中每个事务 t 做以下操作：析出 t 中的频繁项，降序排列。令排列后的序列为 $[p \mid P]$，其中，p 是序列中的第 1 项，P 是余下的序列。调用函数 insert_tree($[p \mid P]$,T)。

函数 insert_tree($[p \mid P]$,T) 的执行过程是：如果 T 有一个子节点 N，使得 N.itemname=p.itemname，那么将 N 节点的计数加 1；否则，为 T 建立一个新子节点 N，令 N.itemname=p，N.count=1，并将 N 链接到头表中同 itemname 链的尾部。如果 P 非空，再将 P 分为 $[p \mid P]$，递归调用函数 insert_tree($[p \mid P]$,T)。

图 2.7　FP-Tree 的构造算法

输入：T　　//根据事务数据库生成的 FP-Tree

　　min_sup　　//最小支持度

输出：全部频繁项集

方法：call FP-growth (T, null)

　FP-growth (Tree, α)

　　if Tree 包含一个单一路径的前缀　then

　　{令 P=Tree 的单一路径前缀部分；

　　将 Tree 去掉 P 后的部分加上一个根 null，赋给 Q；

　　for each P 中节点的组合 β　do

　　产生模式 $\beta \cup \alpha$，令 $\beta \cup \alpha$ 的支持数=β 中支持数最小的项目的支持数

　　将这样产生的模式的集合称为 freq_pattern_set (P);}

　　else 将 Tree 赋给 Q；

　　for each item a_i in Q do

　　{产生模式 $\beta=a_i \cup \alpha$，令 $\beta=a_i \cup \alpha$ 的支持数=a_i.count；

　　构造 β 的条件模式基及其条件模式树 Tree_β；

　　if $\text{Tree}_\beta \neq \varnothing$ then call FP-growth (Tree_β, β)；

　　将这样产生的模式的集合称为 freq_pattern_set (Q);}

　　return {freq_pattern_set(P) ∪ freq_pattern_set(Q) ∪ [freq_pattern_set(P)×freq_pattern_set(Q)]}

图 2.8　FP-growth 算法

2.3 聚类、分类与模式识别

2.3.1 基本概念

聚类是对物理的或抽象的样本集合分组的过程。聚类的目标是把一个样本集合分割为子集或簇，使得簇内部的样本之间的相关性比与其他簇中样本之间的相关性更紧密。聚类的方法主要有划分方法、基于密度的方法、基于网格的方法、层次方法等。聚类是无监督的模式识别的主要手段，也是数据准备过程中连续数据离散化的主要方法。

分类是数据挖掘的一项重要任务，其目标是从已知类标号的训练集中学习模型，并用该模型对类标号未知的记录进行分类。针对分类问题，人们开发了很多算法，较经典的有神经网络方法、支持向量机方法、关联规则分类算法、K 近邻分类算法、决策树分类算法和贝叶斯分类算法等[25,90]。

模式识别的方法在一定程度上可以说是数据挖掘算法的应用或延伸。模式识别是一门以应用为基础的学科，其目的是利用计算机实现人的类识别能力。模式是指具有某种特定性质的观察对象。对象与应用领域有关，它们可以是图像、信号波形或者任何可测量且需要分类的对象[90]。模式类是通过特征来表示的，特征选择的好坏，直接影响分类器的性质。在模式识别系统设计中，特征的确定往往是一个反复的过程，其选择和提取方法对领域知识有较强的依赖性。如果有一个可用的训练数据集，并通过挖掘先验已知信息来设计分类器，则称为有监督模式识别，其方法与数据挖掘的分类异曲同工；如果模式识别没有已知类别标签的训练数据可供使用，则称为无监督的模式识别，聚类方法是其中最典型的方法[90]。

1. 特征提取与选择

在聚类、分类与模式识别系统中，或者明显地或者隐含地要有特征提取与选择技术环节，通常其处于对象特征数据采集和分类识别两个环节之间，特征提取与选择方法的优劣极大地影响着聚类分类器的设计和性能。由于在很多实际问题中常常不容易找到那些最重要的特征，或受条件限制不能对它们进行测量，这就使特征选择和提取的任务复杂化，成为困难的任务之一。

根据分类对象或目的不同，对象的特征数值化结果有下述 3 种类型[92]：

(1) 物理量。直接反映特征的实际物理或几何意义，如重量、速度和长度等。进行处理分析前需要对这些连续量进行离散化。

(2) 次序量。特征在数值化时，按某种规则确定特征的等级，次序量只反映次序关系。此已为离散数据，如产品的等级、人的学识、技能的等级、病症的级或期等。

(3) 名义量。有些特征本身是非数值的，如男性与女性、事物的状态、种类等，为便于分析而将它们数值化。这些特征的数值指标既无数量含义，也无次序关系，只是用数字代表各种状态。

在特征空间中，如果同类模式相距较近，不同类模式相距较远，分类识别就比较容易正确。因此，在提取实际对象的特征时，要求所提取的特征对不同类的对象差别很大而同类对象差别较小，这将给后继分类识别环节带来很大的方便。但是由于某些原因，提取出的特征不具有这些特性。通常在得到实际对象的若干具体特征之后，再由这些原始特征产生出对分类识别最有效、数目最少的特征，这就是特征提取与选择的任务。在实现上述目标时，往往需要首先制定特征提取与选择的准则，可直接以反映类内与类间距离的函数作为准则，或直接以误判概率最小作为准则，也可以用类别判决函数作为准则，还可以构造与误判概率有关的判据来刻画特征对分类识别的贡献或有效性。在具体实施特征提取与选择时有以下两个基本途径[92]：

(1) 当实际用于分类识别的特征数目 N 给定后，直接从已获得的 M 个特征 x_1，x_2，…，x_M 中选择 N 个，使可分性判据 J 的值满足式 (2.2)，即：

$$J(x_1,x_2,\cdots,x_M)=\max[J(x_{i1},x_{i2},\cdots,x_{iM})] \tag{2.2}$$

即寻找 M 维特征空间中使判据 J 最大的 N 维空间。这类方法称为直接选择法，主要有分支定界法。

(2) 在使判据 J 取最大的目标下，对 M 个原始特征进行变换降维，即对原 M 维特征空间进行坐标变换，再进行直接选择。这类方法称为变换法，主要有基于可分性判据的特征提取选择、基于误判概率的特征提取选择、离散 $K-L$ 变换法及基于决策界的特征选择等。

2. 相似性测度

聚类、分类与识别均需解决对象或模式间的相似度问题，距离是表征相似性的主要特征。对不同领域对象、不同类型的数据，其间距离的函数不同，一般以两个矢量的函数来表达。设矢量 x 与 y 的距离为 $\mathrm{d}(x,y)$，一般地讲，$\mathrm{d}(x,y)$ 应满足[92]：

(1) $\mathrm{d}(x,y)\geqslant 0$，当且仅当 $y=x$ 时，等号成立。

(2) $\mathrm{d}(x,y)=\mathrm{d}(y,x)$。

(3) $\mathrm{d}(x,y)\leqslant \mathrm{d}(x,z)+\mathrm{d}(z,y)$。

距离函数可以有各种形式，以下列出主要的几种，大部分基于实际的分类方法使用欧几里得距离函数[92]。

设，$x=(x_1,x_2,\cdots,x_n)^1$，$y=(y_1,y_2,\cdots,y_n)^1$，则有：

1) 欧几里得距离 (Euclidean)。

$$d(x,y)=\sqrt{\sum_{i=1}^{n}(X_i-Y_i)^2} \tag{2.3}$$

2）绝对值距离（Manhattan）。

$$d(x,y)=\sum_{i=1}^{n}|X_i-Y_i| \tag{2.4}$$

3）切氏距离（Chebyshev）。

$$d(x,y)=\max_i|X_i-Y_i| \tag{2.5}$$

4）明氏距离（Minkowski）。

$$d(x,y)=\sum_{i=1}^{n}[|X_i-Y_i|^m]^{\frac{1}{m}} \tag{2.6}$$

另一类表征距离的测度是以两矢量的方向为基础的。设 $x=(x_1,x_2,\cdots,x_n)^1$，$y=(y_1,y_2,\cdots,y_n)^1$，则[91,92]：

1）角度相似系数（夹角余弦）。矢量之间的相似性可用它们的夹角余弦来度量。两个矢量 x 和 y 的夹角余弦如式（2.7）。

$$\cos(x,y)=\frac{X'Y}{[(X'X)(Y'Y)]^{\frac{1}{2}}} \tag{2.7}$$

2）相关系数。它实际上是数据中心化后的矢量夹角余弦。

$$r(x,y)=\frac{(X-\overline{X})'(Y-\overline{Y})}{[(X-\overline{X})'(X-\overline{X})(Y-\overline{Y})'(Y-\overline{Y})]^{\frac{1}{2}}} \tag{2.8}$$

3）指数相似系数。

$$e(x,y)=\frac{1}{n}\sum_{i=1}^{n}\exp\left[-\frac{3}{4}\frac{(X_i-Y_i)^2}{\sigma_1^2}\right] \tag{2.9}$$

式中：σ_1^2 为相应分量的方差。

从函数构造上看，指数相似系数属于距离方式，但从测度值和相似性关系看，属于方向相似测度。

2.3.2 聚类的方法

聚类分析的思想就是根据“物以类聚”的原理，将样本或模式进行分类。聚类的定义与待处理对象特征有关。基于不同的模型构造思想，提出了一系列更加具体化的定义[95]：

（1）基于距离的定义。一个聚类是这样一组数据的集合，其聚类成员之间距离的最大值小于这些成员到任一非聚类成员距离的最小值。显然，这种定义方法过于机械，它对于刻画复杂数据聚类的能力较弱。

（2）基于质心的定义。在数据集合中指定若干个质心，根据各数据点到这些质心的距离划分聚类，它使得一个聚类中的元素到其质心的距离小于其到任

何其他质心的距离。这种定义方法一般只能够恰当地描述球形聚类，它不适于描述不规则形状的聚类。

(3) 基于连接的定义。一个聚类是一组彼此连接的数据点的集合，每个数据到聚类中其他数据之间的接近程度大于其到非聚类成员之间的接近程度。这种定义方法的有效性在于它能够描述任意形状的聚类，但当数据集包含噪声时，基于连接的定义方法难以区分夹杂于噪声数据之间的不同聚类。

(4) 基于密度的定义。聚类是数据空间中的稠密区域元素所构成的子集合，两个聚类之间的边界通过稀疏区域划定。基于密度的定义能够适应聚类形状不规则和包含噪声数据的情形，因此它较之其他定义方法具有独到的优越性。

(5) 基于相似性的定义。聚类是一组“相似”数据对象的集合，处于不同聚类的元素彼此互不“相似”。

聚类的方法很多，不同的方法对于同一数据集聚类的结果可能不同。以下介绍划分法和层次方法[95]。

1. 划分法

给定包含 n 个数据对象的数据集，并给定所要形成的聚类个数 K($K<n$)，划分算法将对象集合划分为 K 份，其中每个划分均代表一个聚类，所形成的聚类将使得同类中的对象是“相似”的，而不同聚类中的对象是“不相似”的。

K-均值 (K-means) 聚类方法是最典型的划分方法，将 n 个对象划分成 K 个聚类，确保聚类内具有较高的相似度，而聚类间的相似度较低。其处理流程为：首先从 n 个数据对象中随机选择 K 个数据对象作为初始聚类的中心，而对于余下的每个对象，根据其与各个聚类中心的距离或相似度，分别将它们分配给与其最相似的聚类；然后重新计算每个聚类的平均值。这个过程不断重复，直到目标测度函数开始收敛。其目标测度函数通常采用平方误差准则，即：

$$E=\sum_{i=1}^{K}\sum_{p\in Ci}|P-m_i|^2 \tag{2.10}$$

式中：E 为所有聚类对象的平方误差和；p 为聚类对象；m_i 为类 C_i 的各聚类对象的平均值。

K-均值算法具体描述如图 2.9 所示。

当结果簇是密集的，而簇与簇之间区别明显时，K-means 算法效果较好。对处理大数据集，该算法是相对可伸缩的和高效率的，因为它的复杂度是 O(nkt)，n 是所有对象的数目，k 是簇的数目，t 是迭代的次数。通常情况下，$k\ll n$，且 $t\ll n$。

输入：聚类的数目 K 和包含 n 个数据对象的数据库，最小平方误差 c

输出：K 个聚类

算法：

(1) 随机选择 K 个对象作为初始的聚类中心。

(2) 将每个对象分给初始的聚类。

(3) Repeat

根据聚类中数据对象的平均值，将每个数据对象重新赋给最相似的聚类。

计算每个聚类中数据对象的平均值，更新聚类的平均值。

(4) Until 平均误差 $E<c$ 或者每个聚类不再发生变化。

图 2.9 K-均值算法

但是，K-means 算法只有在簇的平均值被定义的情况下才能使用，且用户指定的 k 值在很多应用中依据不容易把握。另外，K-means 算法不适合于发现非凸形状的簇，或者大小判别很大的簇。而且，它对于噪声和孤立点敏感，少量的该类数据能够对所属簇的平均值产生极大的影响[25]。

K-means 算法有很多变种。它们可能在初始 k 个平均值的选择，相异度的计算，计算聚类平均值的策略上有所不同。经常会产生较好的聚类结果的一个策略是，首先采用层次的自底向上算法决定的结果簇的数目及找到初始的簇，然后用迭代的重定位来改进聚类结果[25]。

K-means 算法的一个变体是 k-modes 方法，它扩展了 K-means 算法，用模式来替代类的平均值，采用新的相异性度量方法来处理分类性质的数据，采用基于频率的方法来修改聚类的模式。K-means 算法和 k-modes 方法可以综合起来处理有数值类型和分类类型属性的数据，这就是 k-prototypes 方法[25]。

期望最大（Expectation Maximization，EM）算法以另一种方式对 K-means 算法进行了扩展。它不把对象分配给一个确定的簇，而是根据对象与簇之间隶属关系发生的概率来分派对象。换句话说，在簇之间没有严格的界限[25]。

2. *层次方法*

层次方法就是通过分解所给定的数据对象集来创建一个层次。根据层次分解形式的方式，层次聚类方法可划分为凝聚的层次聚类和分裂的层次聚类方法。

(1) 凝聚的层次聚类是自底向上的策略。首先将每个对象作为一个簇，然后合并这些原子簇为越来越大簇，直到所有的对象都在一个簇中，或者某个终结条件被满足。绝大多数层次聚类方法属于这一类，它们的不同表现在簇内与簇间相似度的定义不同。

（2）分裂的层次聚类是自顶向下的策略。首先将所有对象归为一个簇中，然后逐渐细分为越来越小的簇，直到每个对象自成一簇，或者达到了某个终结条件，例如达到了某个希望的簇数目，或者两个最近的簇之间的距离超过了某个阈值。

在凝聚或分裂的层次聚类方法过程中，簇间距离是重要的依据，目前广泛采用的簇间距离的度量方法一般有 4 种[25]：

1）最小距离：$d_{\min}(Ci,Cj)=\min_{p\in Ci,p^1\in Cj}|p-p^1|$。

2）最大距离：$d_{\max}(Ci,Cj)=\max_{p\in Ci,p^1\in Cj}|p-p^1|$。

3）平均值距离：$d_{\text{mean}}(Ci,Cj)=|m_i-m_j|$。

4）平均距离：$d_{\text{avg}}(Ci,Cj)=\sum_{p\in Ci}\sum_{p\in Cj}|p-p^1|$。

这里，$|p-p^1|$是两个对象 p 和 p^1 之间的距离；Ci 是簇的平均值。

由于划分的不可逆性，这种方法的最大困难在于聚类过程中对聚类进行合并或分裂的选择，不适宜地选择合并或分裂会导致低质量的聚类结果。此外，每次决定聚类的合并还是分裂都需要检验和计算大量的数据对象和聚类，因此效率较低[95]。目前，一般将基于层次的聚类方法和其他聚类技术进行集成以形成多阶段聚类，从而提高聚类质量。常用的层次方法有最短距离法、最长距离法、中间距离法、BIRCH、CURE 和 Chameleon 等[95]。

BIRCH（Balanced Iterative Reducing and Clustering Using Hierarchies）是利用层次方法的平衡迭代约减和聚类方法[25]。它引入了两个概念：聚类特征和聚类特征树（CF tree），用于概括聚类过程的描述。它是一种结构辅助聚类的方法，在大型数据库中的速度和伸缩性较好。

一个聚类特征（CF）是一个三元组，用来描述子类（cluster）对象的信息。假设某个子类中有 n 个 d 维的对象或点 $\{o_i\}$，则式（2.11）、式（2.12）、式（2.13）定义了该子类的质心 x_0、半径 R 和直径 D。

$$x_0=\frac{\sum_{i=1}^{n}x_i}{n} \tag{2.11}$$

$$R=\sqrt{\frac{\sum_{i=1}^{n}(x_i-x_0)^2}{n}} \tag{2.12}$$

$$D=\sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i-x_j)^2}{n(n-1)}} \tag{2.13}$$

R 和 D 反映了整个子类围绕它的质心的紧密程度。

而该子类的 CF 定义如式（2.14）。

$$CF=(n,LS,SS) \tag{2.14}$$

式中：n 为子类中点的数目；LS 为 n 个点的和$\left(\sum_{i=1}^{n}x_i\right)$；$SS$ 为 n 个点的平方和$\left(\sum_{i=1}^{n}x_i^2\right)$。

CF 有效地记录了子类的要素，而不是存储所有子类中的对象。

CF 树是高度平衡的树，它存储了层次聚类的特征。如图 2.10 所示是一个 CF 树的示意图（原图来源于文献 [25] 中图 7.8）。树中的非叶节点有后代或子树，非叶节点存储了其后代的 CF 的和，也就是概括了其子节点的聚类信息。一棵 CF 树有两个参数：分支因子 B 和阈值 T。分支因子定义了树中非叶子节点的最大子节点数目，阈值 T 给出了树中叶子节点的最大直径。两参数会影响生成树的大小。

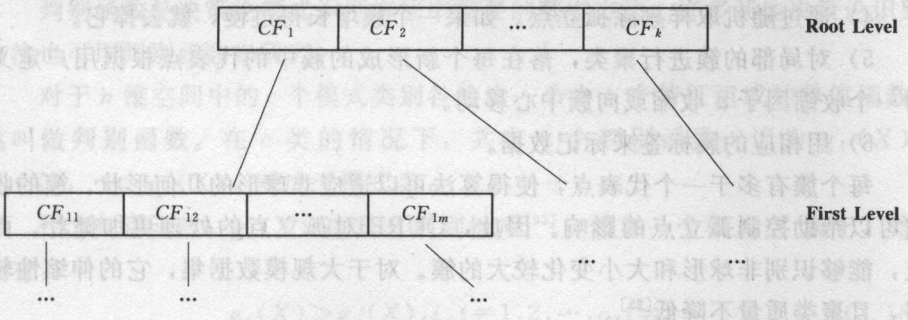


图 2.10　CF 树结构示意图

BIRCH 算法包括两个阶段[25]：

(1) 扫描数据集，建立一个初始 CF 树，CF 树一般存放在内存中。这个 CF 树压缩了数据集信息，只保留了聚类所需的数据信息。

(2) 采用某个聚类算法对 CF 树的节点进行聚类。

在阶段 (1)，随着新的数据对象被处理，CF 树被动态构造，这个方法对增量聚类有很好的支持。一个数据对象被插入到最近的叶子。如果在插入后存储在叶子节点中的子类的直径大于阈值 T，那么该叶子就会被分裂。新的数据对象插入后，其信息向树根传递。如果 CF 树构造过大，不适应内存，可以定义一个较小的阈值，并重建 CF 树。重建过程从旧树的叶子节点构造新树，不需重读数据集。因此，为了构建 CF 树，一般只需读一个数据集。

在阶段 (2)，利用构建的 CF 树来生成最好的聚类。BIRCH 采用了一种多阶段聚类技术，一遍扫描 CF 树产生基本的聚类，再进行一或多遍扫描可以进一步改进聚类的质量。算法的计算复杂度是 O(n)。实验表示 BIRCH 算法对数据集规模伸缩有线性的时间效率，且聚类质量较好。但是，由于 CF 树节

点阈值限制，其聚类结果并不总是表达了自然聚类。而且，如果簇不是球形的，算法不会很好地工作，因为它用半径和直径的概念来控制聚类过程。

大多数聚类算法或者擅长处理球形和相似大小的聚类，或者在存在孤立点时变得比较脆弱。CURE 解决了偏于球形和相似大小的问题，在处理孤立点上也更加健壮。CURE 采用一种新的层次聚类算法，算法选择基于质心和基于代表对象方法之间的中间策略[25]。在 CURE 算法中，一个簇的代表点通过如下方式产生：首先选择簇中分散的对象，然后根据一个特定的分数或收缩因子向簇中心收缩。有算法的每一步，有最近距离的代表点对的两个簇会被合并。CURE 算法的核心步骤描述如下[25]：

1）从源数据对象中抽取一个随机样本 S。

2）将样本 S 划分为一组分块。

3）对每个划分局部聚类。

4）通过随机取样剔除孤立点。如果一个簇增长得过慢，就去掉它。

5）对局部的簇进行聚类，落在每个新形成的簇中的代表点根据用户定义的一个收缩因子 a 收缩或向簇中心移动。

6）用相应的簇标签来标记数据。

每个簇有多于一个代表点，使得算法可以适应非球形的几何形状。簇的收缩可以帮助控制孤立点的影响。因此，CURE 对孤立点的处理更加健壮。而且，能够识别非球形和大小变化较大的簇。对于大规模数据集，它的伸缩性较好，且聚类质量不降低[25]。

2.3.3 K-近邻分类方法

最初的近邻法是由 Cover 和 Hart 于 1968 年提出的，由于对该方法在理论上进行了深入分析，直至现在仍是模式识别非参数法中最重要的方法之一[95]。

假定有 c 个类别 ω_1，ω_2，…，ω_c，的模式识别问题。每个类有标明类别的样本 N_i 个。如果设定 ω_i 类的判别函数为：

$$g_i(X)=\min_i \| X-X_i^k \| ,k=1,2,\cdots,N_i \tag{2.15}$$

式中：X_i^k 的角标 i 表示 ω_i 类；k 表示 ω_i 类 N_i 个样本中的第 k 个。

按照式（2.15），决策规则可以写为：

若 $g_i(X)=\min_i g_i(X)$，则决策 X 属于类 ω_i。

这一决策方法称为最近邻法。就是说对未知样本 x，只要比较 x 与 c 已知类别的样本之间的距离，并决策 X 为与它最近的样本同类[95]。当最近邻法应用于特定的一组样本时，所得到的错误率与样本的偶然有关。

K-近邻法是最近邻法的一个显然的推广。取未知样本 x 的 K 个近邻，看这 K 个近邻中多数属于哪一类，就把 x 归为哪一类。

无论是最近邻法还是K-近邻法，它们都有方法简单的优点，而且其错误率在贝叶斯错误率和两倍贝叶斯错误率之间。正是近邻法的这种优良性质，使它成为模式分类的重要方法之一。但近邻法存在下列问题[95]：

(1) 需将所有样本存入计算机中，每次决策都要计算待识别样本 x 与全部训练样本之间的距离并进行比较，使存储量和计算量都很大。

(2) 虽然在所有情况下，对未知样本 x 都可以进行决策，但当错误代价很大时，会产生较大的风险。

(3) 要求样本趋于无穷大，这在任何实际场合都是无法实现的。

快速搜索近邻法考虑将样本分级分成一些不相交的子集，并在子集的基础上进行搜索。该算法对最近邻法和K-近邻法都适用。

2.3.4 线性判别式分类方法

判别函数是决策论模式识别方法中重要的基本方法，许多其他决策论识别方法也可用判别函数来研究。

对于 n 维空间中的 c 个模式类别各给出一个由 n 个特征组成的单值函数，这叫做判别函数。在 c 类的情况下，共有 c 个判别函数，记为 $g_1(X)$，$g_2(X)$，…，$g_c(X)$，对应于模式类 ω_1，ω_2，…，ω_c。

作为判别函数，$g_i(X)$ 应具有判别性质[95]。假如一个模式 X 属于 i 类，则有：

$$g_i(X)>g_j(X),i,j=1,2,\cdots,c,i\neq j \tag{2.16}$$

而如果这个模式在第 i 类和第 j 类的分界面上，则有：

$$g_i(X)=g_j(X),i,j=1,2,\cdots,c,i\neq j \tag{2.17}$$

最简单的判别函数是线性判别函数，它是所有模式特征的线性组合。对于第 i 类模式，它可表达为：

$$g_i(X)=\lambda_{i1}x_1+\lambda_{i2}x_2+\cdots+\lambda_{in}x_n+\lambda_{i0},i=1,2,\cdots,c \tag{2.18}$$

λ_{ij} 是特征的系数，称为权。式子 (2.18) 说明，每个特征对于判别函数作出不同的贡献，权可以看做其贡献的大小。

所谓设计线件分类器，就是利用训练样本集建立线性判别函数式。建立判别函数式的过程，实际上就是寻找最好权值的过程。如前所述，最好的结果往往出现在准则函数的极值点上，设计线性分类器的问题就转化为利用训练样本集寻找准则函数极值的问题。设计线性分类器的主要步骤可概括如下[96]：

(1) 要有一组具有类别标志的样本集。

(2) 要根据实际情况确定一个准则函数 J。

(3) 用最优化技术求出准则函数的极值解。

这样就获得了每个类别的线性判别函数，对于未知类别的样本 Y，只要计

算 $g_i(Y)$，然后根据 $g_i(Y)$ 值和式（2.16）就可以判断 Y 所属的类别。

2.3.5　支持向量机分类方法

由线性判别函数的设计过程可知，对于线性可分集，总能找到使模式样本正确划分的解。一般说来，它有无穷多个解，希望找到一个最优的解。一种最优的分界准则是使两类模式向量分开的间隔最大。如图 2.11 所示是两类线性可分集被超平面隔开的情况（来源于文献［96］中图 2.14）。图 2.11（b）比图 2.11（a）中两类样本分开的间隔要大，如果图 2.11（b）中的 H_0 是具有最大间隔的解平面，则称这个解平面是最优的，其中距离最优分界面最近的位于间隔边界上的那些模式向量就叫做支持向量。

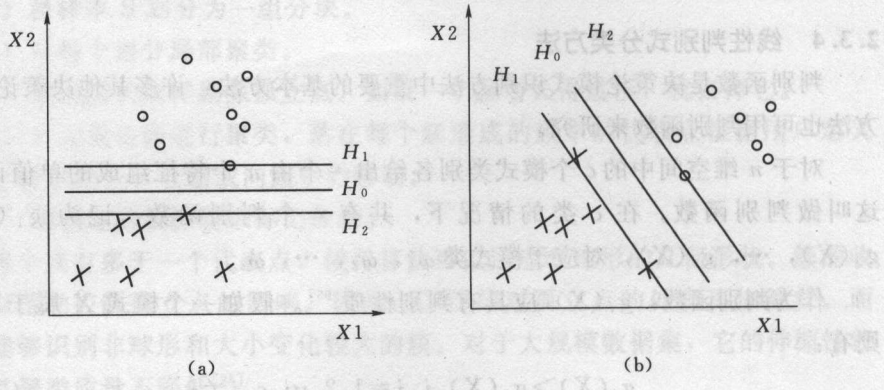


图 2.11　两类线性可分集最优划分示意图

设线性可分两类样本集为 (X_i, Y_i)，$i=1,\cdots,n$，X 为特征属性，Y 值为 $+1$ 或 -1，表示两个类别。d 维空间中线性判别函数的一般形式为 $g(x)=\omega X+b$，分类面方程为：

$$\omega X+b=0 \tag{2.19}$$

将判别函数进行归一化，使两类所有样本都满足 $|g(x)|>1$，这样分类间隔就等于 $2/\|\boldsymbol{\omega}\|$。满足分类条件且 $\|\boldsymbol{\omega}\|$ 最小的分类面就是最优分类面，最终转化为求极值解问题。

对于线性不可分问题，可以用类似于广义线性判别函数的方法，通过事先选择好的非线性映射将输入模式向量映射到一个高维空间，在这个空间中构造最优分界超平面。

2.3.6　神经网络分类方法

1943 年，心理学家 Mcculloch 和数学家 Pitts 合作提出形式神经元的数学模型，成为人工神经网络研究的开端。1949 年，心理学家 Hebb 提出神经元之间突触联系强度可变的假设，并据此提出神经元的学习准则，为神经网络的

学习算法奠定了基础。1982 年，Hopfield 提出了神经网络的一种数学模型，引入了能量函数的概念，研究了网络的动力学性质，又设计出用电子线路实现这一网络的方案，同时开拓了神经网络用于联想记忆和优化计算的新途径，大大促进了神经网络的研究。

神经网络的原始模型是 1958 年提出的感知器模型，只有一个输出节点，它相当于单个神经元，主要用于模式分类。感知器的功能是有限的，它无法解决线性不可分的两类样本的分类问题。之后，人们又提出了多层前馈网络及反向传播训练算法，简称 BP 网络或 BP 算法。

多层前馈网络结构如图 2.12 所示（来源于文献［93］中图 11.3）。构成前馈网络的各神经元接受前一级输入，并输出到下一级、无反馈，形成一有向无环图。除输入和输出层，中间的层称为隐层。图的节点分为两类，即输入节点与计算单元。几乎所有神经网络学习算法都可以看做 Hebb 学习规则的变形。Hebb 学习规则的基本思想是：如果神经元 U_i 接收来自另一神经元 U_j 的输出，则当这两个神经元同时兴奋时，则从 U_j 到 U_i 的权就得到加强。

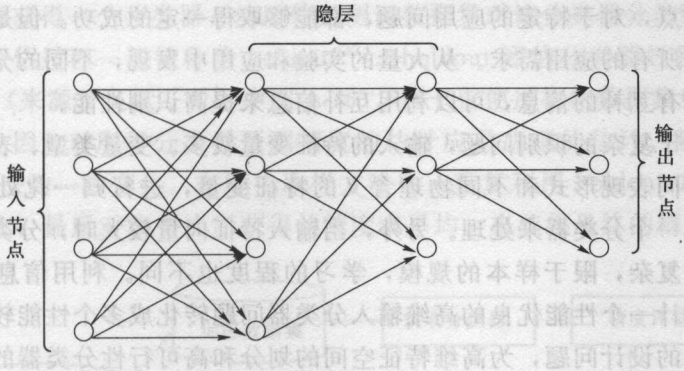


图 2.12　多层前馈网络结构示意图

多层前馈网络的适用范围大大超过原始网络，但其主要困难是中间的隐层不直接与外界连接，无法直接计算其误差。为解决这一问题，提出了反向传播算法。其主要思想是从后向前（反向）逐层传播输出层的误差，以间接算出隐层误差。算法分为两个阶段：第一阶段（正向过程）输入信息从输入层经隐层逐层计算各单元的输出值；第二阶段（反向传播过程）输出误差逐层向前算出隐层各单元的误差，并用此误差修正前层权值。反向传播算法解决了隐层权值修正问题，但它是用梯度法求非线性函数极值，因而有可能陷入局部极小点，不能保证收敛到全局极小点[93]。

利用神经网络进行分类时，如果训练样本是 M 维模式向量，它们依次送到神经网络的输入层，用来训练神经网络的参数。输入层各节点以一定权值送

到第一隐层各节点，第一隐层各节点输出再以一定权值送到第二隐层各节点。最后，第二隐层节点分别送到输出层，得到运算结果。

观察第一隐层诸神经元，它们有共同的输入和不同的输出。每个这样的神经元都是感知器，它决定了多维空间的不同超平面。每个超平面将 M 维空间分割成正负两个半空间。给出的超平面通过以后的连接和组合构成一个正确的分段线性划分，这个分段线性划分能拟合对这些类模式正确分类的超曲面。这就把模式识别中的分段线性划分与前馈神经网络联系了起来[96]。

2.3.7　多分类器融合与分类算法性能评估

1. 多分类器融合

过去 10 多年中，对分类器研究的焦点从单个分类器的研究转移到多分类系统的研究。多分类器融合就是融合多个分类器提供的信息，得到更加精确的分类结果。融合多分类器的特长是系统研究的目标，其必要性体现在两个方面[91]：

(1) 分类方法有很多，这些方法是基于不同的理论框架提出来的。每种分类器各有优点，对于特定的应用问题，都能够取得一定的成功。但是，没有一种方法适应所有的应用需求。从大量的实验和应用中发现，不同的分类器对于分类模式具有互补的信息，可以利用互补信息来提高识别性能。

(2) 对于复杂的识别问题，输入的特征变量较多。变量类型、表现形式不同。对于不同表现形式和不同物理含义的特征变量，进行归一化处理非常困难，难以用一个分类器来处理。另外，当输入特征向量较大时，分类器的结构将变得非常复杂，限于样本的规模，学习的程度也不同。利用信息融合的思想，可能设计一个性能优良的高维输入分类器问题转化成多个性能较优的低维输入分类器的设计问题，为高维特征空间的划分和高可行性分类器的设计提供一个新的思路。

根据各分类器提供的信息的级别，多分类器的融合可分为以下 3 种类型[91]：

(1) 决策层融合，即单个分类器输出为某个确定的类标号。假设 R 个分类器 e_1，e_2，…，e_R 对同一输入标本 x 进行分类，事件 $e_k(x)=j_k$ 表示分类器 e_k 把 x 划分到 Wj_k 中，其中，$j_k\in\Lambda\cup\{M+1\}$，$\{M+1\}$ 表示分类器 e_k 拒识 x，$k=1,2,\cdots,R$。决策层融合就是利用这些事件构造一个集成的分类器 e 对 x 进行分类，输出一个确定的类别号，$e(x)=j$，$j\in\Lambda\cup\{M+1\}$。多数据投票法和 BKS 方法均是决策层的多分类器融合方法。

(2) 排序层融合，即单个分类器输出为样本属于各类的可能性的一个排序列表。对于输入 x，每个分类器 $e_k(x)$ 产生一个子集 $L_k\subseteq\Lambda$，且 L_k 中标签排

列成一个序列。排序层融合就是利用事件 $e_k(x)=L_k, k=1,2,\cdots,R$，构造一个集成的分类器 e，对 x 进行分类，输出一个确定的类别号，$e(x)=j$，$j\in\Lambda\cup\{M+1\}$。

(3) 度量层融合，即单个分类器输出为样本属于相应类的程度。对于输入 x，每个分类 e_k (x) 器产生一个度量向量 $\boldsymbol{M}_e(k)=[m_k(1),m_k(2),\cdots,m_k(M)]^{\mathrm{T}}$，其中，$m_k(i)$ 表示 s 属于相应类 wi 的程度。度量层融合就是利用事件 $e_k(x)=M_e(k)$，$k=1,2,\cdots,R$，构造一个集成的分类器 e，对 x 进行分类，输出一个确定的类别号，$e(x)=j, j\in\Lambda\cup\{M+1\}$。

3 类方法利用的分类器输出信息量集资增多，相应地也可能得到更好的结果。

2. 分类算法性能评估

分类精度是度量分类效果的最主要指标，一般定义为被分类器正确分类的对象数与被分类的所有对象数之比。为测量分类精度，一般将数据集随机划分为两个部分：一个作为训练数据集；另一个则作为测试数据。通常训练集包含初始数据集的 2/3 的数据，而其余的 1/3 则作为测试数据集的内容。利用训练集数据学习获得一个分类器，然后使用测试数据集对该分类器分类精度进行评估。这种评估过程一般被称为 holdout[97]。holdout 评估分类器过程示意如图 2.13 所示（来源于文献 [97] 中图 4.16）。由于仅使用初始数据集中的一部分进行学习，因此对所得分类器预测精度的估计应是悲观的估计。随机取样是 holdout 方法的一种变化。在随机取样方法中，重复利用 holdout 方法进行精度评估 k 次，最后对这 k 次所获得的精度求平均，来获得最终的精度值。

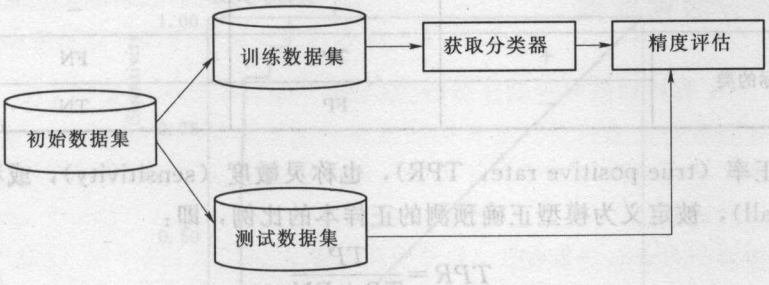


图 2.13 用 holdout 方法评估分类器精度的过程

k-交叉检验（k-cross validation）是随机二次抽样方法的交叉验证。初始数据集被随机分为 k 个互不相交的子集 S_1，S_2，…，S_k，每个子集大小基本相同。学习和测试过程分别进行 k 次，每一次选择一个子集作为测试集，其他子集则合并到一起构成一个大训练数据集，通过学习获得分类器，对测试集进行测试，获得精度。一般第 1 次选择 S_1 作为测试集，第 2 次选择 S_2 作为

测试集……最终的精度可用 k 次测试的平均精度来表征。

除精度外，评估分类算法性能的还包括其速度、鲁棒性、可规模性和可解释性等。鲁棒性一般评估分类器对噪声的敏感程度；可规模性可以通过计算给定分类算法在渐增的数据集上的 I/O 操作次数评估；可解释性是主观的度量。

对于不平衡数据集的二分类问题，数据集中只有大类和小类两类，由于两类数据比例不均衡，精度值一般不能准确表征分类的效果。例如，对于网络访问数据集，其记录一般 90%以上是大类数据，是正常访问的数据，小于 10%的记录是入侵访问。如果简单地将所有记录判定成正常记录，其正常记录的分类精度也会大于 90%，而这个值是无意义的。在这类数据集中，二类记录不是同等重要的，受关注的是小类记录，一般称为正类样本，而大类记录称为负类样本。为此，一些新的度量指标是必需的[97]。为说明这些指标，需声明 4 个数值术语的定义：真正、假负、假正、真负[97]。混淆矩阵是用来表示这 4 个术语的汇总表格，其示意如表 2.2 所示（源于文献［97］中的表 5.6）。

1）真正（true positive，TP）对应于被分类模型正确预测的正样本数。

2）假负（false negative，FN）对应于被分类模型错误预测为负类的正样本数。

3）假正（false positive，FP）对应于被分类模型错误预测为正类的负样本数。

4）真负（true negative，TN）对应于被分类模型正确预测的负样本数。

表 2.2　类不同等重要的二分类问题混淆矩阵

		预测的类	
		+	−
实际的类	+	TP	FN
	−	FP	TN

真正率（true positive rate，TPR），也称灵敏度（sensitivity），或称召回率（recall），被定义为模型正确预测的正样本的比例，即：

$$TPR=\frac{TP}{TP+FN} \tag{2.20}$$

真负率（true negative rate，TNR），也称特指度（specificity），定义为被分类器正确预测的负样本的比例，即：

$$TNR=\frac{TN}{TN+FP} \tag{2.21}$$

假正率（false positive rate，FPR），定义为被分类器预测为正类的负样本比例，即：

$$FPR=\frac{FP}{TN+FP} \tag{2.22}$$

假负率（false negative rate，FNR），定义为被分类器预测为负类的正样本比例，即：

$$FNR=\frac{FN}{TP+FN} \tag{2.23}$$

那么，前述的精度即被正确分类的正样本数与被正确分类的负样本数之和占全部样本数的比例，表达为：

$$\text{accuracy}=\frac{TP+TN}{TP+FN+FP+TN} \tag{2.24}$$

可以构造一个基线模型，它最大化其中一个度量而不管另一个。例如，将每个记录都声明为正类，则召回率永远是1，但它的精度等值会很差。构建一个最大化精度和召回率的模型是分类算法的一个主要挑战。

另外，一个可以调和精度和召回率的指标 $F1$ 被定义为：

$$F1=\frac{2TP}{2TP+FN+FP} \tag{2.25}$$

接受者操作特征曲线（receiver operating characteristic curve，ROC）是显示分类器真正率和假正率之间折中的一种图形化方法[97]。在ROC图中，一般 x 轴表示假正率（在SPSS分析结果图中用1-sepecificity标记），y 轴表示真正率（也称灵敏度，在SPSS的分析结果图中用sensitivity标记），其示意如图2.14所示［来源于本书第3.4节的图3.6（a）］。

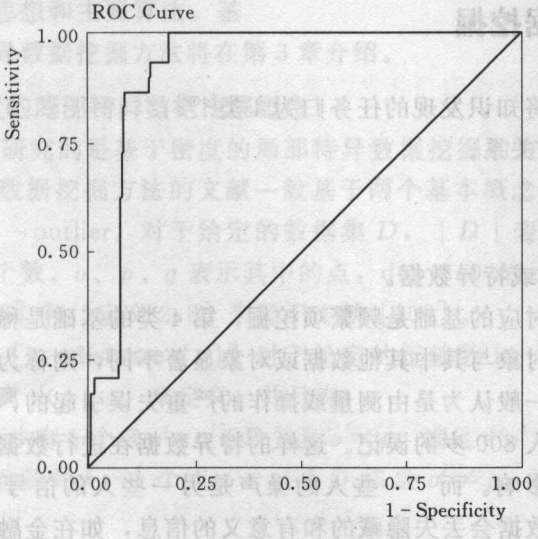


图2.14 ROC曲线示例

ROC 曲线上有几个关键点：

1）（$TPR=0$，$FPR=0$）：把每个实例都预测为负类的模型。

2）（$TPR=1$，$FPR=1$）：把每个实例都预测为正的模型。

3）（$TPR=1$，$FPR=0$）：理想模型。

如果一个 ROC 曲线越靠图的左上角，则分类器的性能就越好。主对角线上对应的分类器应是一个随机分类的模型。

ROC 曲线下方的面积提供了评价分类器平均性能的一种方法。理想模型 ROC 曲线下方的面积等于 1。随机分类模型 ROC 曲线下方的面积等于 0.5。如果一个分类器 ROC 曲线下方面积大于另一个分类器 ROC 曲线下方面积，则其性能好于另一个分类器。

欲绘制 ROC 曲线，分类器应当能够产生可以用来评价它预测的连续值输出，从最有可能成为正类的记录到最不可能成为正类的记录进行排序。利用排序后的序列绘制 ROC 曲线的过程一般为[97]：

1）选择正类可能性最高的记录，把选择的记录和可能性低于它的记录指定为负类，相当于把所有记录均指定为负类。那么，所有负类均被正确分类，所有正类均被错误分类，$TPR=FPR=0$。

2）从排序序列中取下一记录，把选择的记录和可能性低于它的记录指定为负类，相当于只把第一个记录指定为正类，计算此时的 TPR 和 FPR 的值。

3）重复步骤 2），并计算 TPR 和 FPR 的值，直到序列中所有的值计算完。

4）根据产生的 TPR 和 FPR 的值绘制 ROC 曲线。

2.4　特异数据挖掘

E. Knorr 等将知识发现的任务归为 4 类[109]：

1）探查依赖关系。

2）辨识类别。

3）描述类别。

4）探查例外或特异数据。

前 3 类任务对应的基础是频繁项挖掘，第 4 类的基础是稀有项。在数据集中，一些数据或对象与其中其他数据或对象显著不同，则称为特异数据或特异对象。特异数据一般认为是由测量或操作的严重失误引起的，如在登记年龄的数据中出现了某人 800 岁的误记。这样的特异数据在进行数据分析前是应尽量清除掉或减少其影响。而“一些人的噪声是另一些人的信号”[108]，在另一些场合，清除特异数据会丢失隐藏的和有意义的信息，如在金融领域使用信用卡的特异消费可能是欺诈行为。在这些应用中，发现特异数据成为挖掘的目标。

Muneaki Ohshima 和 Ning Zhong 等人认为数据（属性值）只被少数对象拥有，并且与其他数据显著不同，则这些数据是特异的（peculiar）[101]。Hawkins 将特异数据定义为："如果一个值与其他值差距很大，以至让人怀疑它是由不同的机制产生的，这个值是特异值（outlier）"[105]。虽然没有统一的概念，但特异数据被公认有两个特征：稀少和与其他数据的差距大。

典型的特异数据挖掘算法有以下几种：

1）基于统计的。

2）基于密度的局部特异数据挖掘方法。

3）基于距离的全局特异数据挖掘方法。

另外，还有一些文献研究基于数据挖掘中间结果的特异识别，如利用分类关联规则来识别特异的类别等[113]。其中，基于统计的方法，主要是利用数据的分布特性计算特异数据的特征，采用不一致检验的方法挖掘数据。因为现实数据的分布特性往往不是已知的，而根据数据来计算分布特性是相当复杂的；并且，有些现实的数据集没有一致的分布状态，例如如图 2.15 所示的数据集。所以基于统计的方法的应用很受限制。2）和 3）的方法均从数据本身出发挖掘特异数据，本章将介绍基于密度的局部特异数据挖掘方法的思想和主要算法。基于距离的全局特异数据挖掘方法将在第 3 章介绍。

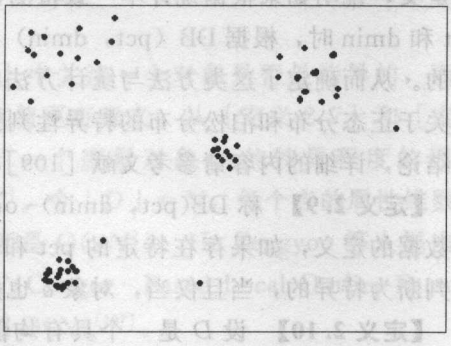


图 2.15 无一致分布数据集示例

2.4.1 基于密度的局部特异数据挖掘概念

文献［106］研究的是基于密度的局部特异数据挖掘的方法和理论。基于密度的局部特异数据挖掘方法的文献一般基于两个基本概念：k－distance 和 DB（pct，dmin）－outlier。对于给定的数据集 D，$|D|$ 表示其中点（或对象、或记录）的个数。o、p、q 表示其中的点，d(p，o) 表示 p 和 o 的距离。下面的［定义 2.7］和［定义 2.8］来源于文献［106］。

【定义 2.7】［k－distance（p）］ 对于给定的正整数 k，k－distance（p）用点 p 和 o 的距离 d(p，o) 来定义，并且有：

1）在 D 中至少有 k 个点 o'，$o' \in D$ 并且 $o' \neq p$，满足 d(o'，p)$\leqslant$d(o，p)。

2）在 D 中至多有 $k-1$ 个点 o'，$o' \in D$ 并且 $o' \neq p$，满足 d(o'，p)$<$d(o，p)。

显然，k－distance(p) 越大，p 点附近的点密度越低，p 的特异程度越高。

【定义 2.8】 **[DB (pct, dmin)- outlier]** 称一个点 p 是 DB(pct, dmin)- outlier，如果在 D 中至少有百分之 pct 的点与 p 之间的距离大于指定的最小距离 dmin。也就是集合 $\{q \in D \mid d(p, q) \leqslant \text{dmin}\}$ 的势$\leqslant (100-\text{pct})/100 \times |D|$。

例如：对给定的 D，设定 pct=98.5，dmin=10。如果存在点 p，使得 D 中至少 98.5%的点与 p 的距离大于 10，那么称 p 是 DB(98.5, 10)- outlier。从另一个角度讲，如果 p 是 DB(98.5, 10)- outlier，那么 D 中至多有 1.5% 的点与 p 的距离$\leqslant$10。

由定义可知，可以通过 k - distance(p) 来确定 DB(pct, dmin)- outlier。因为 k - distance(p) 只取决于 p 附近点的值，所以基于此两个概念开发的算法被归类为局部的方法。

E. Knorr 和 R. Ng 等在文献 [109] 中详细分析了 DB(pct, dmin)- outlier 的定义，说明如果根据统计不一致检测，对象 o 是特异的，那么当给定适当的 pct 和 dmin 时，根据 DB (pct, dmin) - outlier 定义也会将对象 o 判定为是特异的。从而确定了这类方法与统计方法的一致性。以下只列出文献 [109] 中的关于正态分布和泊松分布的特异性判别与 DB - outlier 判别关系的相关定义与结论，详细的内容请参考文献 [109] 原文。

【定义 2.9】 称 DB(pct, dmin)- outlier 的定义统一了或概括了另一个特异数据的定义，如果存在特定的 pct 和 dmin，使得依据另一个定义，对象 o 被判断为特异的，当且仅当，对象 o 也是 DB - outlier。

【定义 2.10】 设 D 是一个具有均值 μ 和标准差 σ 的正态分布的对象集，p 是 D 中的一对象。正态分布的特异定义 $\text{Def}_{\text{Normal}}$：$p$ 是特异的，当且仅当，$\dfrac{p-\mu}{\sigma} \geqslant 3$ 或 $\dfrac{p-\mu}{\sigma} \leqslant -3$。

此定义将正态分布中与均值距离不小于 3 的对象称为是特异的。

【定理 2.1】 定义 DB(pct, dmin)- outlier 统一了定义 $\text{Def}_{\text{Normal}}$，当 pct=0.9988，dmin=0.13$\sigma$ 时，也就是依据 $\text{Def}_{\text{Normal}}$，对象 p 是特异的，当且仅当，p 是 DB(0.9988, 0.13σ)- outlier。

【定义 2.11】 说明了当参数 μ=3 时，泊松分布的特异数据界定。

【定义 2.12】 泊松分布的特异数据定义 $\text{Def}_{\text{Poisson}}$ 为：p 是特异的，当且仅当，$p \geqslant 8$。

【定义 2.13】 DB(0.9962, 1)- outlier 统一了 $\text{Def}_{\text{Poisson}}$。

2.4.2　基于密度的局部特异数据挖掘方法

E. Knorr 和 R. Ng 等在文献 [109] 中同时提出了一种基于网格构架的挖掘 DB - outlier 的方法。首先，全部的数据空间被分割成边长为 l 的网格。称在 x 行和 y 列的单元格为 $C_{x,y}$，依据式 (2.26) 和式 (2.27) 定义其 L_1 邻居

和 L_2 邻居。

$$L_1(C_{x,y})=\{C_{u,v}\mid u=x\pm1,v=y\pm1,C_{u,v}\neq C_{x,y}\} \tag{2.26}$$

$$L_2(C_{x,y})=\{C_{u,v}\mid u=x\pm3,v=y\pm3,C_{u,v}\neq C_{x,y}\} \tag{2.27}$$

根据［定义 2.7］，给定一正整数 k，E. Knorr 和 R. Ng 等提出的算法根据［性质 2.3］编制。

【性质 2.3】 ①如果 $C_{x,y}$ 中的对象数$>k$，那么 $C_{x,y}$ 中的所有对象均不是特异对象。②如果 $C_{x,y}\cup L_1(C_{x,y})$ 中的对象数$>k$，那么 $C_{x,y}$ 中的所有对象均不是特异对象。③如果 $C_{x,y}\cup L_1(C_{x,y})\cup L_2(C_{x,y})$ 中的对象数$\leqslant k$，那么每一个 $C_{x,y}\cup L_1(C_{x,y})\cup L_2(C_{x,y})$ 中对象均是特异对象。

可以看出，此方法将对象 VS 对象的处理过程简化为单元- by -单元的处理过程，从而获得了时间效率。令 $|D|=N$，每个点的属性维数为 α，此方法总的时间复杂度为 $O(c^2+N)$。其中，c 是一个与维数及每维划分的格数有关的常数。

M. M. Breuning 等认为文献［106］中关注一个对象是否是特异的，而很多的应用中，给出一个对象的特异程度值更有意义。从［定义 2.7］和［定义 2.8］出发，M. M. Breuning 等定义了一个度量对象 p 的特异程度的因子 Local Outlier Factor，记为 LOF(p)[106]。令 $|D|=N$，每个点的属性维数为 α，计算每个点 LOF(p) 的时间复杂度是 $O(\alpha N^2)$。He Zengyou 等人提出了一种基于聚类的方法，提出了 CBLOF(Cluster - Based Local Outlier Factor) 因子及计算算法，其时间复杂度降低到 $O(N)$[107]。

2.5 数据挖掘应用现状

近 30 年，数据挖掘技术已取得了重大的进展。其技术已应用在广泛的领域，很多商业的数据挖掘系统已可用，但是数据挖掘技术与应用方向还面临很多挑战[25]。

很多银行和金融机构均提供广泛的信用服务、投资服务和保险服务等。银行和金融机构积累的数据全面、可信，质量很高，基于此的数据分析非常有效。数据仓库、数据立方体、数据提取和分类，以及特异分析在这类部门中发挥了重要作用。这方面典型的应用还包括：还贷预测和客户信用评估、为目标业务的客户分类与聚类、洗钱和其他金融犯罪的探察等[25]。

在生物医药领域，功能性染色体、蛋白质组成和生物制药的研究在近 10 年中进展迅猛。DNA 序列构成了所有生命基因代码的基础，氨基酸是构成蛋白质的基石，其中包含着生命特征的决定信息，分析其结构序列是生物研究的重要手段。数据挖掘技术在生物领域研究中扮演着重要角色，如多蛋白质序列

中的连接、检索、相似性搜索和对比分析；基因网络和蛋白质路径的结构分析；共生基因序列的关联分析等[25]。

另外，数据挖掘技术还广泛应用在零售业的销售、客户、产品、季节及地区分析，促销业绩分析、客户忠诚度分析、产品推荐分析等；应用在计算机安全领域的入侵检测；应用在从 Web 上发现行业情报以及在视频录像中发现特异行为等。理论上讲，在需要从大量的数据中发现知识和规律的场合均有数据挖掘的用武之地。但是，由于各领域知识背景复杂，数据的复杂性、动态性，数据挖掘理论技术与领域知识技术的融合还面临诸多挑战，激励人们做深入探索。

第3章　基于聚类的全局特异数据挖掘算法

特异数据挖掘是数据挖掘的重要研究方向，其概念和方法的研究贯穿整个数据挖掘研究历程。在一些现实问题中特异数据挖掘具有举足轻重的意义。如计算机安全领域的网络入侵识别、在金融领域的信用卡欺诈甄别和洗钱行为判断、医学领域的疾病诊断等。这些问题中要识别的对象或数据没有固定的属性特征，但它们有共同的特点："特异"。第2.4节中介绍了基于密度的局部特异数据挖掘概念与方法，本章将阐述一类基于距离的全局特异数据挖掘的概念和方法，提出一种基于聚类的全局特异数据挖掘算法，并通过实验检测算法的效率。

3.1　基于距离的全局特异数据挖掘概念和方法

Zhong Ning 等人提出了一种基于距离的全局特异数据挖掘的构架[101]。以下将概略阐述此构架的思想和基本概念，细节内容请参考文献［101］。

首先，特异数据挖掘有4个基本问题要考虑：

1）对于一个关系数据集，特异分析有两个层次：属性层次和记录层次。

2）对于多个关系集，要考虑它们的主、外键关系。

3）要选择合适的方式来表达结果模式。

4）要考虑算法的效率。

给定一个数据集 D，D 是由属性和记录组成的二维关系表。对其中特异数据的挖掘分为两个层次：挖掘每个属性中的特异数据（Attribute - level）和挖掘特异记录（Record - level）。设数据集 D 有 n 个记录 $\{X_1, X_2, \cdots, X_n\}$、$m$ 个属性 $\{a_1, a_2, \cdots, a_m\}$，其结构如表3.1所示。

表3.1　　数据集 D 结构示意表

	a_1	a_2	…	a_j	…	a_m
X_1	x_{11}	x_{12}	…	x_{1j}	…	x_{1m}
X_2	x_{21}	x_{22}	…	x_{2j}	…	x_{2m}
⋮	⋮	⋮	…	⋮	…	⋮
X_i	x_{i1}	x_{i2}	…	x_{ij}	…	x_{im}
⋮	⋮	⋮	…	⋮	…	⋮
X_n	x_{n1}	x_{n2}	…	x_{nj}	…	x_{nm}

对属性层次，表3.1中的任一属性 x_{ij}，其特异因子（peculiar factor）记为 $PF(x_{ij})$，用式（3.1）来计算。式中，$d(x_{ij}-x_{kj})$ 表示属性 x_{ij} 和 x_{kj} 间的距离，α 默认值为0.5，可根据需要调节。

$$PF(x_{ij})=\sum_{k=1}^{n} d(x_{ij}-x_{kj})^{\alpha} \tag{3.1}$$

一个阈值 p_j 由式（3.2）来定义。式中，γ 的默认值为1，可根据需要调节。其中，M_j 和 σ_j 分别代表 a_j 所有 PF 值的均值和标准差。当 $PF(x_{ij})\geqslant p_j$ 时，称 x_{ij} 是特异的。

$$P_j=M_j+\gamma\times\sigma_j \tag{3.2}$$

对记录层次，记录 X_i 的特异因子记为 $RPF(X_i)$ 由式（3.3）计算。式中，β_j 是属性 a_j 的权，默认值全部取1。

$$RPF(X_i)=\sum_{k=1}^{n}\sqrt{\sum_{j=1}^{m}\beta_j\left[PF(x_{ij})-PF(x_{kj})\right]^2} \tag{3.3}$$

一个阈值 pr 由式（3.4）定义。式中 γ 的取值类似于式（3.2）。当 $RPF(X_i)\geqslant Pr$ 时，称 X_i 是特异的。

$$Pr=RPF\text{ 均值}+\gamma\times RPF\text{ 的标准差} \tag{3.4}$$

可以看出，此构架是基于距离的，并且找到的特异数据是全局的。从式（3.1）得出，属性 x_{ij} 和 x_{kj} 间的距离 $d(x_{ij}-x_{kj})$ 是后续计算的基础。对不同特征的属性，距离的算法不同，说明如下：

（1）当 a_j 是非主键属性时

1）当 a_j 是数值型属性，并且没有其他的背景知识可用时，a_j 中不同值的距离可由式（3.5）计算：

$$d(x_{ij}-x_{kj})=|x_{ij}-x_{kj}| \tag{3.5}$$

由式（3.1）、式（3.3）和式（3.5）可以看出，不同数值属性的计量单位会对 RPF 有不均衡的影响。为避免这些影响，可以将属性值先进行标准化，再计算距离。一种简单的标准化方法是将属性值转化成无纲量的变量。设 x_{1j}，…，x_{nj} 是属性 a_j 的全部得值。M_j 是它们的平均值，σ_j 是它们的标准差。那么，可以用式（3.6）将每个值进行标准化。

$$x_{ij}'=\frac{x_{ij}-m_j}{\sigma_j} \tag{3.6}$$

那么式（3.5）就可转化为：

$$d(x_{ij}-x_{kj})=|x_{ij}'-x_{kj}'| \tag{3.7}$$

对于属性层次，是否将属性值进行了标准化并不影响 PF 值的序列关系，对应值是否是特异的判断也就没有影响。基于此，也可以在计算出 PF 后，将 PF 标准化，来代替属性层次的标准化。这样做的好处是可以不考虑属性的类

型。PF 标准化的公式可采用式（3.8）。其中，M_j 和 σ_j 的意义与式（3.2）中相同。而在式（3.3）中利用标准化后的 PF 值（称 PF'）计算 RPF。

$$PF'(x_{ij})=\frac{|PF(x_{ij})-M_j|}{\sigma_j} \tag{3.8}$$

2）当 a_j 是符号型属性，并且没有其他的背景知识可用时，简单地设置相同符号间距离为 0，不同符号间距离为 1。

3）当 a_j 是符号型属性，并且有背景知识可用时，按背景知识计算属性值间的距离。

4）当 a_j 是日期型时，可以用两日期值的间隔作为其距离。

（2）当 a_j 是主键或外键属性时

1）如果 a_j 是数据集 D 的主键，那么它只是每个记录的标识，对计算属性或记录的特异性没有意义。

2）如果 a_j 是数据集 D 的外键，那么它是另一数据集的主键，可以把它在主键数据集中的 RPF 值作为在本数据集中的属性值。

由前述可知，此构架可以处理各类属性值，并且可以处理多个相关数据集的特异挖掘问题。给定一个数据集 D，可以计算每个属性的特异因子以及每个记录的特异因子，并且可根据阈值 p 和 pr 判断它们的特异性。将找出一个数据集 D 中特异的属性和特异记录的过程总结为算法，为叙述方便，称其为 PecuFind 算法，如图 3.1 所示。PecuFind 算法采用的是标准化 PF 的方法，采用属性值标准化的算法可以依此类推。

输入：D—数据集，有 m 个属性，n 个记录
输出：属性 PF 和记录 RPF
算法：
（1）根据式（3.1）计算属性特异因子 $PF(x_{ij})$，$i=1, 2, \cdots, n$；$j=1, 2, \cdots, m$。
（2）计算 PF 的均值 M_j 和标准差 σ_j，$j=1, 2, \cdots, m$。
（3）根据式（3.2）计算属性特异阈值 p_j，$j=1, 2, \cdots, m$。
（4）根据式（3.8）将属性 PF 标准化。
（5）根据式（3.3）计算记录特异因子 $RPF(X_i)$，$i=1, 2, \cdots, n$。
（6）根据式（3.4）计算记录特异阈值 pr。
（7）排序输出 PF 和 RPF，并根据 p 和 pr 标注其特异性。

图 3.1 PecuFind 算法

3.2 一种基于聚类的全局特异数据挖掘算法

特异数据在整个数据集中占的比例很小，第 3.1 节的构架中有大量的冗余运算。设某一属性的所有值的数据集为 S，其平均值为 S_{mean}。如果事先根据

距离将 S 进行聚类，那么大类（其中数据个数多的类）中的数值成为特异的可能性就会很小，质心（类中数据的平均值）距 S_{mean} 较近的类中的数值成为特异的可能性也会很小。第3.1节的构架精确地计算出了所有属性与记录的特异因子，而用户关心的只是特异值排序在前面的很少一部分属性与记录的特异因子值，对于排序在后面的特异因子值可以采用不精确的计算方法。根据这些想法，提出一种基于聚类的全局特异数据挖掘方法。

给定一个数据集 D，D 是由属性和记录组成的二维关系表。数据集的结构如表3.1所示。首先，一个百分数 λ 需要由用户给出。λ 表示数据集 D 中特异数据占的比例，可以是用户预测 D 中有 $\leqslant \lambda \times |D|$ 个特异数据或用户希望能找出前 $\lambda \times |D|$ 个特异值。另外，还需要设置一个聚类系数 k，其值为2或3或4，可由用户指定，默认 $k=2$。k 的意义将在后面说明。构架仍由挖掘特异属性和挖掘特异记录两个层次构成。

在属性层次，首先，将每个属性按距离进行聚类；之后，计算每个属性聚类的反特异因子，并将属性聚类分成可能成为特异的类和不可能成为特异的类；最后，计算属性的特异因子并排序，指定前 $\lambda \times |D|$ 属性为特异的。

设某一属性的所有值的集合为 S，其平均值为 S_{mean}。从原则上讲可以采用任何基于距离的聚类算法对 S 进行聚类，采用的聚类算法的效果好，可以减少后续的计算量。为了时间效率，使用的是一种类似于 Squeezer 算法[111]和文献［112］中聚类算法的简单聚类算法，称 SimC（Simple Cluster），如图3.2所示。SimC 算法思想简单，运行效率高。虽然聚类的结果并不优秀，但后面的实验将证明，SimC 有效地支持了本章的构架。

输入：S—数据集
　　λ—用户指定的百分数
　　k—默认值为2，可由用户调整为3或4
输出：聚类结果集合 $\{C_1, C_2, \cdots\}$
算法：
　（1）找出 S 中的最大值和最小值 S_{max}，S_{min}。
　（2）计算聚类半径 $Cd=(S_{max}-S_{min})\times\lambda/k$。
　（3）$\eta=1$。
　（4）Val=S 的第1个值，生成一个新类 C_1，Val 加入类 C_1，令 C_1 类的均值 C_{1mean}=Val。
　（5）对 S 中剩余的每个数据，做：
　1）Val=S 中的一个值。
　2）计算 Val 与 $C_1 \sim C_\eta$ 的均值的差，找到其中最小的差 d_j。
　3）如果 $d_j \leqslant Cd$，将 Val 加入类 C_j 中，重新计算均值 C_{jmean}。
　否则，$\eta=\eta+1$，生成一个新类 C_η，Val 加入 C_η，令 C_η 的均值 $C_{\eta mean}$=Val。
　（6）输出 C_1，C_2，…，C_η。

图3.2　SimC 聚类算法

可以看出，k 是控制聚类半径 Cd 的。k 越大，Cd 越小，产生的类越多，同时每个类更精细。设 S 经聚类生成 $\{C_1, C_2, \cdots, C_\eta\}$ η 个类。现在根据式（3.9）计算每个类的特异因子，记为 $CPF(C_j)$。式中，$|C_j|$ 表示类 C_j 中元素的个数。

$$CPF(C_j)=\frac{|S_{\text{mean}}-C_{j\text{mean}}|}{\sqrt{|C_j|}} \tag{3.9}$$

显然，CPF 越小的类，其中的元素是特异数据的可能性越小。将 η 个类按 CPF 由大到小的顺序排序，设排序后的顺序为：$C'_1, C'_2, \cdots, C'_\eta$。再设置一个阈值，称红蓝界 RBB，并且 $RBB=|S|\times\lambda\times\left(1+\frac{1}{\sqrt{k}}\right)$，$k$ 是之前设定的聚类系数。从 C'_1 开始累计类中元素个数，设当累计到 C'_{jr} 时结果第 1 次大于或等于 RBB，则称 $C'_1, C'_2, \cdots, C'_{jr}$ 为红类；$C'_{jr+1}, C'_{jr+2}, \cdots, C'_\eta$ 为蓝类。

计算每个属性的特异因子 PF 时，红类中的元素将作为单个的值参加运算，也会有自己的 PF 值，蓝类将整个类参加运算，类中的元素共用一个 PF 值。设经聚类和分类后，类的序列为：$C'_1, C'_2, \cdots, C'_{jr}, C'_{jr+1}, C'_{jr+2}, \cdots, C'_\eta$，且红类为前 jr 个，所有红类中元素总个数的和为 nr，属性 PF 值新的计算方法如式（3.10）～式（3.12）。

$$PF(x_{ij})=\sum_{k=1}^{nr}(X_{ij}-X_{kj})^\alpha-\sum_{k=jr+1}^{\eta}[(X_{ij}-C'_{\text{kmean}})^\alpha\times|C'_k|] \tag{3.10}$$

$$x_{ij},x_{kj}\in\text{红类},C'_k\in\text{蓝类}$$

$$PF(x_{ij})=PF(C'_j) \qquad x_{ij}\in C'_j,C'_j\text{属于蓝类} \tag{3.11}$$

$$PF(C'_j)=\sum_{k=1}^{\eta}[(C'_{j\text{mean}}-C'_{\text{kmean}})^\alpha\times|C'_k|] \quad C'_j\text{属于蓝类} \tag{3.12}$$

由式（3.10）～式（3.12）可以计算出表 3.1 中全部属性的 PF 值。计算记录的 RPF 时，对式（3.3）稍作变动，变为式（3.13），从先横向再纵向变到先纵向再横向。因为有大批属蓝类的属性有相同的 PF 值，式（3.3）的计算就可大大简化。

$$RPF(X_i)=\sum_{j=1}^{m}\{\beta_j\times\sqrt{\sum_{k=1}^{n}[PF(x_{ij})-PF(x_{kj})]^2}\} \tag{3.13}$$

从聚类开始，经过前面的计算，可以得到每个属性的 PF 值和每个记录的 RPF 值。与第 3.1 节的构架不同的是，这里不需要 p 和 pr 来判断属性和记录的特异性，因为用户指定了 λ，所以将 PF 从大到小排序，RPF 从大到小排序，分别指定前 $\lambda\times|D|$ 个为特异的即可。

对于二维关系表的属性 $a_1, a_2, \cdots, a_m$ 的处理，与第 3.1 节类似。如果 a_j 是表的主键，它是用来标识记录的，不需要计算其 PF 值；如果 a_j 是表的

外键，那么它是另一个表的主键，其 PF 值由那张表的 RPF 来确定；只有 a_j 是表的一般属性时，才有必要计算其 PF 值。

对于不能区分等级的文本属性值，在聚类时可以简单地将具有相同值的聚为一类；计算类的 CPF 时，使用公式 $\frac{1}{\sqrt{|C_j|}}$；计算两个值的距离时，按相同为0，不同为1的原则进行；计算属性 PF 时，只使用式（3.12）按类计算，其余的计算均可照常进行。

在使用式（3.13）计算整个记录的 $RPF(X_i)$ 时，用户可以调节 β_j 来控制各属性的权。为避免属性自身值的不平衡而给 RPF 造成的影响，计算 RPF 前应对属性的 PF 进行标准化（当然也可以采用将属性标准化的方法）。因为计算 PF 的标准差 σ 的时间复杂度是 $O(n^2)$，可以采用式（3.14）来进行 PF 的标准化。其中，$PF_{j\max}$ 和 $PF_{j\min}$ 分别是属性 a_j 的 PF 值中的最大值和最小值。

$$PF'(x_{ij})=\frac{PF(x_{ij})-PF_{j\min}}{PF_{j\max}-PF_{j\min}} \tag{3.14}$$

根据新的构架，将特异计算过程总结为算法，称为 CpecuFind 算法，如图3.3所示。

输入：D—数据集，有 m 个属性，n 个记录

λ—用户指定的百分数

k—默认值为2，可由用户调整

输出：属性 PF 和记录 RPF

算法：

（1）利用 SimC 聚类算法将每个属性分别聚类；

（2）计算各个类的 CNPF，并做红蓝划分；

（3）根据式（3.10），式（3.11）和式（3.12）计算属性特异因子 $PF(x_{ij})$，并排序，$i=1,2,\cdots,n$，$j=1,2,\cdots,m$；

（4）根据式（3.14）将属性 PF 标准化；

（5）根据式（3.13）计算记录特异因子 $RPF(X_i)$，并排序，$i=1,2,\cdots,n$；

（6）输出 PF 和 RPF，并标注前 $\lambda\times|D|$ 个为特异。

图3.3 CpecuFind 算法

3.3 挖掘特异数据能力实验分析

为了验证 CpecuFind 算法挖掘特异数据能力，分别进行了3组实验对比分析。实验机器为 1.8GCPU/512M 内存的笔记本电脑。首先，使用一个小数据

集-人口数据集，测试 PecuFind 算法和 CpecuFind 算法的运行结果，来验证 CpecuFind 算法的有效性。其次，采用 KDDCUP99 数据集，渐增数据集规模，对比 PecuFind 算法和 CpecuFind 算法的结果，验证 CpecuFind 算法挖掘特异数据的能力。最后，采用 Wisconsin Breast Cancer 数据集，并将数据集进行与文献［107］中相同的处理，CpecuFind 算法运行结果与 CpecuFind 算法运行结果对比，及与文献［107］的实验结果进行对比，验证 CpecuFind 算法挖掘特异数据能力，对比局部方法与全局方法挖掘特异数据的能力。

3.3.1 人口数据集实验

人口数据集来自 2006 年的《中国统计年鉴》，表名称是“各地区人口数和出生率、死亡率、自然增长率（2005 年）”。表中共 31 个记录，有 5 个属性，分别是地区、年底人口数、出生率、死亡率和自然增长率。其中地区是关键字，在计算特异因子时使用了后 4 个属性。如表 3.2 所示中是使用 PecuFind 算法和 CpecuFind 算法对此数据集进行特异挖掘的结果，为节省篇幅，表中只列出了记录特异因子的值（下同）。

表 3.2　PecuFind 算法和 CpecuFind 算法对人口数据集的计算结果

序号	PecuFind 算法（$pr=460.2072$）		CpecuFind 算法（$\lambda=20\%$，$k=2$）	
	地区	*RPF* 值	地区	*RPF* 值
1	西　藏	603.4405	西　藏	13.31055
2	广　东	599.4615	广　东	12.4421
3	新　疆	518.9425	新　疆	11.16995
4	宁　夏	496.6072	宁　夏	11.14251
5	河　南	478.3825	河　南	10.71839
6	山　东	466.4695	山　东	10.53245
7	贵　州	418.302	辽　宁	9.587289

其中，PecuFind 算法计算出的 $pr=460.2072$，也就是说前 6 个地区会被标识为特异，约占总数的 20%。从现实的数据来看，*RPF* 值高的地区，其数据确有特异之处，例如：西藏地区人口总数最少，人口出生率和自然增长率奇高，广东省、河南省、山东省的人口数最多等。为了与 PecuFind 算法的结果进行对比，CpecuFind 算法中的参数取 $\lambda=20\%$，$k=2$。从表 3.2 可以得出两个结果前 20%（6 个）排序完全相同，其 *RPF* 具体数值的差异主要是因为采用了不同的标准化方法，从而说明 CpecuFind 算法是有效的。

3.3.2 在 KDDCUP99 数据集上的实验

KDDCUP99 数据集是网络访问数据记录集[117]，它包含了若干个数据集，

本书选用的是 corrected. gz。其中的记录有两大类：正常访问记录和网络攻击记录，而网络攻击的记录又分为若干小类，在本实验中将只按两大类来区分记录。每个记录有 42 个属性，前 41 个是访问特征属性，最后一个属性是记录的类别标识。实验中可以选择前 41 个属性来计算特异因子。从 corrected 中按比例分别选择两类记录来构造若干子集，其中攻击记录所占比例均小于 10%，以使其为特异记录。用“攻击记录数＋正常记录数”来表示这些子集的记录规模，如“30＋300”表示整个子集有 330 个记录，其中 30 个攻击记录，300 个正常记录。分别在 30＋300、50＋500、100＋1000、200＋2000、300＋3000 共 5 个子集上分别运行 PecuFind 算法程序和 CpecuFind 算法程序。在 CpecuFind 中，$\lambda=10\%$，$k=2$。两个算法在 30＋300 和 50＋500 上的运行结果中，*RPF* 值排序在前 10% 和 15% 的记录包含的攻击记录数显示于表 3.3。很明显，CpecuFind 发现攻击记录的能力强于 PecuFind。

表 3.3　　30＋300 和 50＋500 上的实验结果比较

数据集 30＋300			数据集 50＋500		
RPF 排序前 n 个（占总记录数的百分比）	含攻击记录数（占全部攻击记录的百分比）		*RPF* 排序前 n 个（占总记录数的百分比）	含攻击记录数（占全部攻击记录的百分比）	
	CpecuFind	PecuFind		CpecuFind	PecuFind
33（10%）	24（80%）	10（33%）	55（10%）	45（90%）	39（78%）
50（15%）	30（100%）	24（80%）	83（15%）	50（100%）	45（90%）

3.3.3　在 Wisconsin Breast Cancer 数据集上的实验

Wisconsin Breast Cancer 数据集有 699 条记录，每条记录都是记录一位胸部有肿瘤的病人的特征数据。其中 458 条是良性的，241 条是恶性的。记录属性有 11 个，其中第 1 个是关键字 ID，最后一个是良性、恶性标识，中间的 9 个是要计算的特征属性。可以从中随机选择 444 条良性记录和 39 条恶性记录组成一数据集。这个数据集与文献［107］的实验数据集相吻合。文献［107］描述的是基于聚类的且基于密度局部特异数据挖掘方法—FindCBLOF，在 Wisconsin Breast Cancer Data 上的实验结果显示了其挖掘特异数据的能力在同类算法中是优越的。

取 $\lambda=10\%$，$k=2$，使用 CpecuFind 程序进行计算，如表 3.4 显示了其结果与 pecuFind 算法挖掘结果，以及与文献［107］的实验结果对比情况。应该说明的是，因为无法获得文献［107］的原数据集，表 3.4 的结果并不具有严格的可比性。结果再次表明 CpecuFind 挖掘特异数据的能力略优于 pecuFind。同时，粗略地说明基于密度的局部方法和基于距离的全局方法在挖掘特异数据的能力上是相当的。

表 3.4 Wisconsin Breast Cancer 数据集上的实验结果比较

特异值排前 n 个记录（占全部记录的百分比）	其中包含恶性记录个数（占全部恶性记录的百分比）		
	CpecuFind	pecuFind	FindCBLOF
4（0.8%）	4（10.3%）	4（10.3%）	4（10.3%）
8（1.7%）	8（20.5%）	8（20.5%）	7（17.9%）
16（3.3%）	16（41%）	15（38.5%）	14（35.9%）
24（5.0%）	22（56.4%）	21（53.8%）	21（53.8%）
32（6.6%）	27（69.2%）	26（66.7%）	27（69.2%）
40（8.3%）	32（82.1%）	32（82.1%）	32（82.1%）
48（9.9%）	36（92.3%）	34（87.2%）	35（89.7%）

3.4 算法性能实验分析

1. 算法时间效率分析

以属性层次为讨论依据，Zhong Ning 的算法时间复杂度明显是 $O(N^2)$，其中 N 是属性集合中的数据数目。对于本书提出的算法，设经聚类后，聚类的个数为 k；经计算聚类的 CPF 后，需以数据单独参加 PF 值计算的数据个数与以类整体参加 PF 值计算的类的个数之和为 n。在聚类阶段，采用的聚类算法时间复杂度是线性的，为 $O(N)$；聚类后计算每个聚类的 CPF 值阶段时间复杂度是 $O(k)$；最后计算 PF 阶段的时间复杂度是 $O(n^2)$。其中，最后阶段的时间复杂度是平方级的，希望 n 与 N 的比值很小，最好与 N 的增长保持线性以下的增长速度。在第 3.3.2 小节的数据集 30＋300、50＋500、100＋1000、200＋2000、300＋3000 上运行 CpecuFind 算法程序，λ 取 10%，采集在运行过程中 n 与 N 的关系，得到图 3.4。其中，A17、A18、A19 3 条曲线显示的是在数据集中第 17、第 18、第 19 个属性集在实验中 n 随 N 变化的情况，作为对比，直线 10%显示的是 n 与 N 比值为 10%的直线。

由图 3.4 可知，n 与 N 的比值会在给定 λ 周围变化。因不同属性的数据集特征不同，聚类结果不同，n 的变化没有统一的规律。由于特异挖掘中 λ 值很小，提出的算法具有较好的可扩展性。

仍采用在第 3.3.2 小节的渐增数据集 30＋300、50＋500、100＋1000、200＋2000、300＋3000，在其上分别运行 PecuFind 算法和 CpecuFind 算法。在 CpecuFind 算法程序中，λ 取 10%。如图 3.5 所示显示出了两个算法程序的运行时间。其中 PecuFind 在运行“100＋1000”子集时，运行时间已超过 1800s（0.5h），运行被中止，图 3.5 中用“*”表示。此图验证了前一自然段

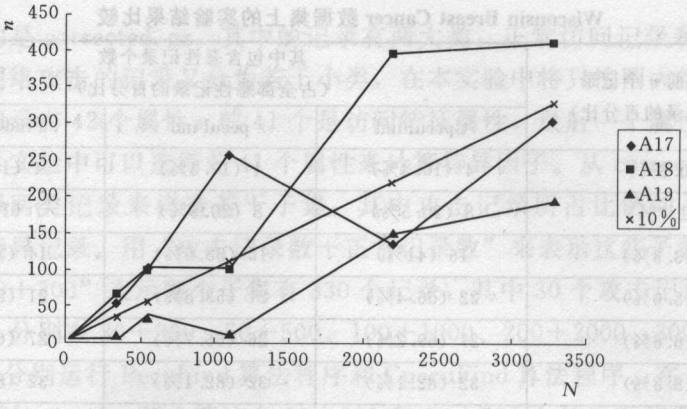


图 3.4　不同数据集规模下 n 与 N 的关系图

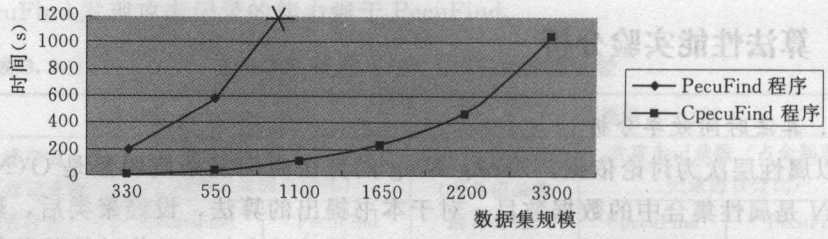


图 3.5　PecuFind 程序和 CpecuFind 程序运行时间比较

中的时间效率分析，表明 CpecuFind 时间效率的优势显著。

2. 两算法 ROC 曲线对比

KDDCUP99 数据集是网络访问数据记录集，其中的记录有两大类：正常访问记录和网络攻击记录，而网络攻击记录在数据集中占的比例很小，如果将攻击记录设为正类，正常访问记录设为负类，本章讨论的特异数据挖掘问题在此数据集上就是不平衡数据集的二分类问题。这类问题的结果评估，除前述的挖掘特异数据能力，还可以以 ROC 曲线特征来评估。对于第 3.3.2 节的 30＋300 和 50＋500 数据集，分别利用 CpecuFind 和 PecuFind 算法计算出的记录特异因子 RPF 值绘制 ROC 曲线，得到 4 个 ROC 曲线图，如图 3.6 和图 3.7 所示，曲线 x 轴表示假正率（Sensitivity），y 轴表示真正率（1－Specificity）。4 个 ROC 曲线下方面积的对比结果如表 3.5 所示。

表 3.5　　4 个 ROC 曲线下方面积对比结果

	PecuFind	CpecuFind
30＋300 数据集	0.924	0.973
50＋500 数据集	0.963	0.987

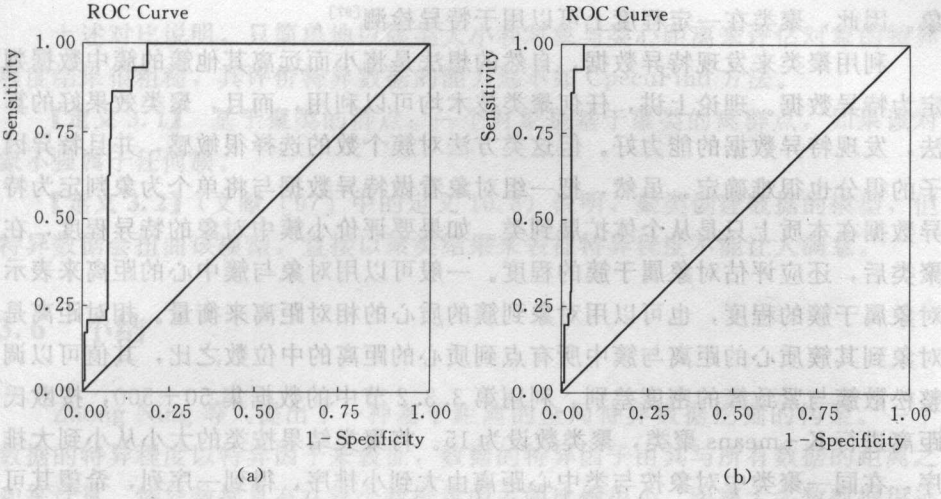


图 3.6 两算法在 30+330 数据集上计算的 ROC 曲线

(a) PecuFind; (b) CpecuFind

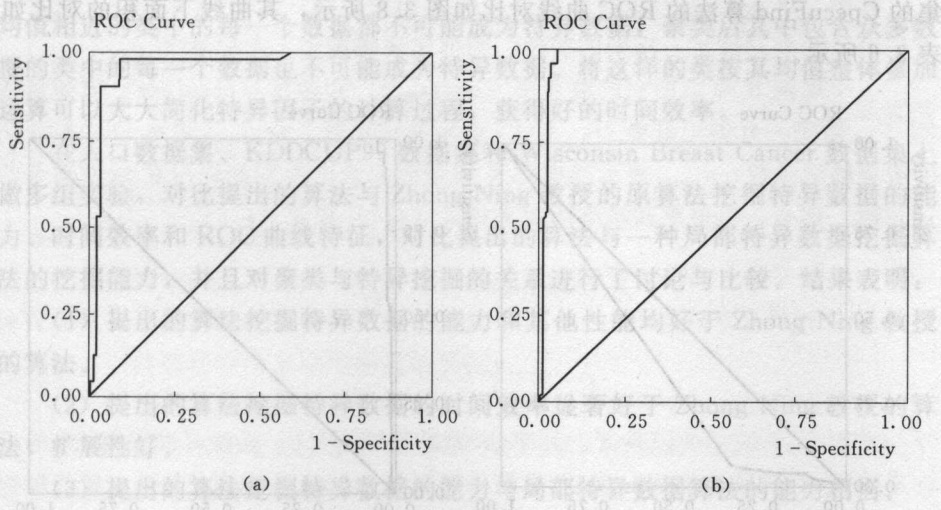


图 3.7 两算法在 550 数据集上计算的 ROC 曲线

(a) PecuFind; (b) CpecuFind

由此，4 个 ROC 曲线的对比说明，CpecuFind 算法性能优于 Pecufind 算法性能。

3.5 聚类算法与特异发现算法对比

聚类分析发现强相关的对象组，而特异检测发现不与其他对象强相关的对

象。因此，聚类在一定程度上可以用于特异检测[97]。

利用聚类来发现特异数据，自然的想法是将小而远离其他簇的簇中数据判定为特异数据。理论上讲，任何聚类技术均可以利用，而且，聚类效果好的算法，发现特异数据的能力好。但这类方法对簇个数的选择很敏感，并且特异因子的得分也很难确定。虽然，把一组对象看做特异数据与将单个为象判定为特异数据在本质上只是从个体扩展到类。如果要评价小簇中对象的特异程度，在聚类后，还应评估对象属于簇的程度。一般可以用对象与簇中心的距离来表示对象属于簇的程度，也可以用对象到簇的质心的相对距离来衡量。相对距离是对象到其簇质心的距离与簇中所有点到质心的距离的中位数之比，其值可以调整松散簇与紧致簇的密度差别。利用第3.3.2节中的数据集50+500，按欧氏距离进行K-means聚类，聚类数设为15。将聚类结果按类的大小从小到大排序，在同一聚类中对象按与类中心距离由大到小排序，得到一序列，希望其可以代表对象的特异程度从大到小的序列。根据对象的真实类别设小类（攻击类）为正类，大类（正常类）为负类，绘制此序列的ROC曲线，其与同数据集的CpecuFind算法的ROC曲线对比如图3.8所示，其曲线下面积的对比如表3.6所示。

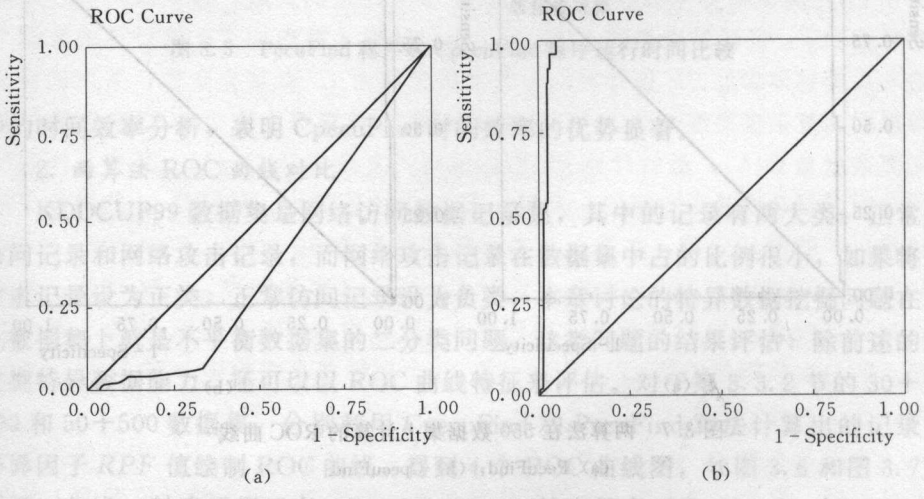


图3.8 K-means聚类方法与CpecuFind发现特异数据能力ROC曲线

(a) K-means聚类；(b) CpecuFind

表3.6 K-means与CpecuFind发现特异数据能力ROC曲线面积对比

	K-means聚类方法	CpecuFind方法
50+500数据集	0.366	0.987

上述对比说明，只简单地以簇类大小和对象与类心距离来评估对象的特异程度结果很粗糙，其评价特异数据的能力远不如 CpecuFind 方法。

【定义 3.1】 基于聚类的群点：一个对象是基于聚类的离群点，如果该对象不强属于任何簇。

【定义 3.2】（文献［97］中的定义 10.8）说明，聚类创建数据的模型，但特异数据会扭曲该模型，直接以聚类结果来表征特异程度不能让人满意。

3.6 小结

Zhong Ning 等人提出了一种基于距离的全局特异数据挖掘的构架。一个数据的特异程度以特异因子来表征，数据的特异因子由其与所有数据的距离之和来计算。特异数据具有在整个数据集中占的比例很小，距离大多数数据很远的特点。由此，提出了先将数据按距离聚类，再根据聚类结果计算数据特异因子的算法。因为要计算的是全局特异因子，聚类后其类的均值与整个数据集的均值相近的类中的每一个数据都不可能成为特异数据，聚类后其中包含众多数据的类中的每一个数据也不可能成为特异数据。将这样的类按其均值整体参加运算可以大大简化特异因子的计算过程，获得好的时间效率。

在人口数据集、KDDCUP99 数据集和 Wisconsin Breast Cancer 数据集上做多组实验，对比提出的算法与 Zhong Ning 教授的原算法挖掘特异数据的能力、时间效率和 ROC 曲线特征，对比提出的算法与一种局部特异数据挖掘算法的挖掘能力，并且对聚类与特异挖掘的关系进行了讨论与比较。结果表明：

（1）提出的算法挖掘特异数据的能力和其他性能均好于 Zhong Ning 教授的算法。

（2）提出的算法挖掘特异数据的时间效率显著好于 Zhong Ning 教授的算法，扩展性好。

（3）提出的算法挖掘特异数据的能力与局部特异数据算法的能力相当。

（4）聚类分析发现强相关的对象组，在一定程度上讲，聚类后小类中的对象特异性更强，可以发现特异数据。但是，在不附加特异因子计算的情况下，聚类方法评价对象特异程度的能力弱于专门的特异数据挖掘方法。

第4章　基于规则的分类方法

4.1　基本概念

规则是表达信息很好的方式。基于规则的分类器利用 IF - THEN 规则集进行分类。如一个规则 $R1$ 可以表示为：

$R1$：IF age＝youth AND student＝yes THEN buys _ computer＝yes

$R1$ 也可以表示为：

$R1$：(age＝youth)^(student＝yes)＝＞(buys_computer＝yes)

其中，IF 部分或＝＞前面的部分称为前件，后一部分称为后件。给定一个记录，如果它满足一规则的前件，则称此规则覆盖了此记录[25]。

一个规则 R 可以用它的覆盖率（coverage）和精度（accuracy）来衡量性能[25]。给定一分类记录集 D，$|D|$ 表示其记录数。令 n_{covers} 表示 D 中被规则 R 覆盖的记录数；n_{correct} 表示被规则 R 覆盖的记录中类别标识与 R 后件吻合的记录数。那么，覆盖率和精度的定义如式（4.1）和式（4.2）。

$$\text{Coverage}(R)=\frac{n_{\text{covers}}}{|D|} \tag{4.1}$$

$$\text{Accuracy}(R)=\frac{n_{\text{correct}}}{n_{\text{covers}}} \tag{4.2}$$

对于 D 中的一个记录 X，如果 X 满足某规则 $R1$，是否可以依据 $R1$ 的后件来判定 X 所属类别呢？当 X 只满足规则集中的一个规则 $R1$ 时，就可以判定 X 属于 $R1$ 后件的类别。但是，当 X 同时满足规则集中的多个规则，并且这些规则的后件不同时，或者当 X 不满足规则集中所有的规则时，X 应该怎样判定呢？

当 X 同时满足规则集中的多个规则时，一般有两种解决方式，称为规格排序方式（size ordering）和规则排序方式（rule ordering）[25]。

所谓规格排序，是在 X 满足的规则中选择条件最强的规则，以此规则的后件来判定 X 的类别。

规则排序方式则预先将所有规则按优先级从高到低排序，形成一个决策列表。X 被判定为在列表中第一次遇到的满足其前件的规则的后件的类别，X

还可能满足的排序在列表后面的规则被忽略掉了。

规则的优先级策略一般有基于类别的（class－based）和基于规则的（rule－based）两种。在基于类别的排序策略中，将规则按其后件类别的重要性从高到低排序，或者将规则按其后件类别误判的代价从高到低排序。后件为同一类别的规则不需排序，因为它们判断的结果不会有冲突；在基于规则的排序策略中，规则按其前件的质量排序。衡量前件的质量有很多种依据，如精度、覆盖率、尺寸（涉及属性的个数）和领域知识等。

当 X 不满足规则集中所有的规则时，一般将 X 指定为一个默认的类别。默认的类别可以是训练集中记录最多的类别，也可以是在训练集中没有被规则覆盖的多数记录的类别。

4.2 基于规则的分类方法

设数据集 D 由记录组成，记录数为 $|D|$，属性有 $\{A_1, A_2, \cdots, A_n, C\}$，其中，$\{A_1, A_2, \cdots, A_n\}$ 是条件属性，C 是类标号。仔细研究各算法就会发现，决策树分类算法、关联规则分类算法、贝叶斯分类算法都是基于规则“$A \rightarrow C$”和其统计特性的。此处，A 表示 $\{A_1, A_2, \cdots, A_n\}$ 全部或部分属性的一些取值组成的集合，C 表示某个类标号。

$C4.5$ 是决策树分类算法的代表[98]。$C4.5$ 方法首先生成一棵未经剪枝的决策树，生成过程遵循如下的原则：

（1）决策树的每个内部节点对应样本的一个非类别属性，该节点的每棵子树代表这个属性的取值范围的一个子区间（子集）。一个叶节点代表从根节点到该叶节点的路径对应的样本所属的类别。

（2）构造决策树时，总选择增益比例大的属性作为下一分支节点。

（3）训练样本集中的未知属性用常用值代替，或者用该属性所有取值的平均值代替，从而处理缺少属性值的训练样本。

将决策树转换为 IF…THEN 规则集合，此集合中规则形如“$A \rightarrow C$”，但其中存在大量的冗余和错误。$C4.5$ 按 K 迭代交叉验证方法选择优化的模型。简化后的规则按类进行分组，形成最终的分类规则集。

朴素贝叶斯分类器是基于贝叶斯概率理论的算法[100]。算法假定在给定类变量的条件下各个属性变量之间条件独立。首先从训练集中计算出每个属性取不同值时各类记录出现的概率，以此作为先验概率。设类标号 $c_j \in C$，分类器利用式（4.3）来判断无类标号记录 $(a_1, a_2, \cdots, a_n)$ 属于类 c_j 的后验概率，并且将此记录判定为属于后验概率最大的类。可见，贝叶斯分类器也是基于规则“$A \rightarrow C$”的统计特性的。

$$P(c_j \mid a_1, a_2, \cdots, a_n) = \frac{P(a_1, a_2, \cdots, a_n \mid c_j) P(c_j)}{P(a_1, a_2, \cdots, a_n)} \tag{4.3}$$

决策树分类法是一种直观且精度较高的方法，但决策树有时也会变得很复杂，以至于难以解释。对于复杂的决策树，从中抽取的IF…THEN规则更加简洁和容易理解。

将决策树的每个分支从根节点开始写成逻辑与的条件，形成IF部分，将叶节点作为THEN部分，就形成了初始的规则集。在此规则集中的规则间隐含着逻辑或的关系，规则间是排他的和无遗漏的，规则间的顺序无关紧要。初始的规则集并不比决策树简单，甚至更难理解，其中可能含有无关和重复的条件，对其进行修剪是必需的[25]。

如何剪枝呢？一般来讲，任何不能提高分类精度的条件均可剪掉。悲观错误策略是常用的策略。其策略是，当欲从规则集中剪掉一个规则中的一个条件或删除一个规则时，判断剪除前和剪除后的错误率，如果错误率没有增加，则实施剪除，否则不剪除。使用这个策略可能使剩余的规则不再排他，也不再无遗漏，使用其进行分类之前应将规则排序，以避免冲突。

使用序列覆盖算法（sequential covering algorithm）可以从训练集中直接抽取IF…THEN规则，不需预先生成决策树[25]。所谓序列覆盖，是指规则的生成过程是每次一条的连续过程。每一条规则都覆盖训练集中的一些记录，希望这些记录均属于一类别。其基本思路是，每一次生成一条新的规则时，被此规则覆盖的记录将被删除，生成和删除过程反复进行，直到满足结束条件。结束条件可能是训练集中没有剩余记录，或者规则的质量已满足用户设置的条件。

一条规则的生成一般遵循由一般到特殊的方式[25]。从前件为空开始，逐步增加前件的条件（属性值），条件之间是“与”的关系。每增加一个条件进行一次测试，满足激励条件的则实施增加，否则不增加。激励条件一般为信息熵降低或信息增益增加等。全部规则生成后，一般还要剪枝，因为生成过程是一个乐观过程，其规则不一定适应后续的测试。

4.3 关联规则分类算法

关联规则是形如“$A \rightarrow C$”的规则，有两个指标：属性集的支持度Support和规则的置信度Confidence。如果后件“C”只有类别值，自然可以想到使用此规则进行分类。典型的关联规则分类算法有CBA、CMAR和CPAR[99]。

CBA是由Liu B等人提出的，是最早的也是最简单的关联规则分类算

法[99]。首先，根据指定的支持度阈值和置信度阈值，在训练集中找出所有形如“$A \rightarrow C$”的关联规则，这类规则被称为“类关联规则”（Class Association Rules，CARs），其特点是后件只包含类标号，这样产生的 CARs 作为初始的规则集。CBA 算法中这部分采用的方法类似于 Apriori。之后进行如下的一系列的操作，以得到最终的规则集。

（1）初始规则集中的规则按置信度、支持度和产生顺序排序，形成一层次结构。

（2）使用类似序列覆盖的方法，将层次规则集作用于训练集，使最终规则覆盖训练集。对于排序在后的规则，如果其没有提高测试精度或者出错率增高了，则剪枝。

（3）在经过剪枝的规则集中加入训练集中的多数类作为默认的类别。

最终规则集就是一张决策表。测试时一个记录被判定为属于在列表中第一次遇到的满足其前件的规则的类别，当此记录不满足所有规则的前件时，判定其为默认类别。在一些典型的分类数据集的测试中，CBA 算法的实验精度高于 $C4.5$。

W Li 等人进一步提出了 CBA 的改进算法 CMAR[114]，其与 CBA 的有 3 点不同之处：

1）在挖掘 CARs 时采用的是 FP－growth 算法。

2）从初始规则集构造最终分类规则集时的策略不同。CMAR 将规则按置信度、相关性和覆盖率进行排序，形成层次结构。每一记录在将插入此结构时，均进行剪枝计算。特殊性更强、置信度更低的规则将被删除。例如，对于规则 $R1$ 和 $R2$，如果 $R1$ 的前件比 $R2$ 的前件更通用，并且 $R1$ 的置信度更大，则 $R2$ 将被剪除。CMAR 还以 X^2 验证测试一规则的前件与后件，当它们不是正相关时，此规则被剪除。

3）分类时，当一记录 X 满足若干规则的前件，CBA 将用置信度最高的规则来判定 X。CMAR 则将这些规则按后件的类别分组，计算各组规则的 X^2 关联强度，用关联最强的组来判定 X。

CMAR 实验精度略高于 CBA，其时间效率、可扩展性和内存利用率均比 CBA 高。

CBA 和 CMAR 均采用频繁项挖掘的方法来产生初始的规则集，根据给定的支持度和置信度，这样产生的规则数据大，在后面要进行大量剪枝工作，从而损失了效率。CPAR 基于 FOIL 算法（一种基于精度的算法）产生规则[115]。对每一类在训练集中找出其正类和反类（其余的类），计算各规则 FOIL 的值，找出得值高的规则加入规则集。同时，覆盖的记录不直接删除，而是降低其权重，以便为其他类的 FOIL 计算提供基数。进行分类时，当一记录满足若干规

则的前件，CPAR 将这些规则按后件的类别分组。但与 CMAR 不同的是，CPAR 只选择每组中的“最好的”K 个规则进行分类计算。CPAR 与 CMAR 分类精度相当，但 CPAR 效率高于 CMAR，特别是在大数据集的分类过程中。

可见，关联规则分类算法是一类基于规则的算法，其基础是频繁项目的与运算。

4.4 必要置信度对分类精度影响的研究

设数据集 D 由记录组成，记录数为 $|D|$，属性有 $\{A1, A2, \cdots, An, C\}$，其中，$\{A1, A2, \cdots, An\}$ 是条件属性，C 是类标号。仔细研究基于规则的算法就会发现，决策树分类算法、关联规则分类算法、贝叶斯分类算法都是基于规则“$A \rightarrow C$”和其统计特性的，此处，A 表示 $\{A1, A2, \cdots, An\}$ 全部或部分属性的一些取值组成的集合，C 表示某个类标号。

在逻辑推理中如果有“$A \rightarrow C$”，则称 A 蕴含 C，或称 A 为 C 的充分条件。对应的，如果有式“$C \rightarrow A$”，则称 A 是 C 的必要条件。由于现实事件很难用全部和精确的数据来描述，并且规则“$A \rightarrow C$”在 D 中的置信度一般不会达到 1，所以在数据挖掘的文献中不称“A 为 C 的充分条件”。但是认为 A 的出现增加了结果为 C 的概率或可能性。可以看出，$C4.5$ 算法、CBA 相关算法和朴素贝叶斯分类器都是基于此原理的。从另一个角度来讲，如果 D 中规则“$C \rightarrow A$”的置信度达到一定的阈值，则可以认为 A 不出现减少了结果为 C 的概率或可能性。在分类算法中如果同时考虑“$C \rightarrow A$”的作用，会提高分类精度。事实是不是这样的呢？本节将通过实验来说明。

4.4.1 问题描述

本节的分类问题描述为：设 $D1$ 和 $D2$ 是针对相同领域的两个数据集，是二维数据表形式，有相同的属性，结构如 $\{A1, A2, \cdots, An, C\}$。$D2$ 中的类标号的集合包含于 $D1$ 中类标号的集合。以 $D1$ 为训练集获取分类规则，用来对 $D2$ 的记录进行分类。

本实验目标是测试两种方法的分类效果：

(1) 方法 1 只考虑规则“$A \rightarrow C$”的影响。

(2) 方法 2 同时考虑“$A \rightarrow C$”和“$C \rightarrow A$”影响。分类效果有很多评价指标，分类精度是重要的指标，其计算公式如式 (4.4)。

$$\text{分类精度} = \frac{\text{被正确分类的记录数}}{\text{待分类的记录总数}} \tag{4.4}$$

基于规则的形式，将形如“$A \rightarrow C$”的规则称为充分规则，形如“$C \rightarrow A$”的规则称为必要规则。为了叙述方便，将“$A \rightarrow C$”的置信度 Confidence 称为

充分置信度，将“$C \rightarrow A$”的置信度称为必要置信度，记为 $N\_Confidence$，并由式（4.5）来计算。

$$N\_Confidence = \frac{D \text{ 中同时包含 } A \text{ 和 } C \text{ 的记录数}}{D \text{ 中包含 } C \text{ 的记录数}} \tag{4.5}$$

4.4.2 实验方法

本实验目标是测试必要规则对分类的影响，将采用最简单的方法。训练过程只采集各属性不同取值与各类之间规则的置信度，不采集组合属性与类之间规则的置信度。当然，现实数据各属性之间一般有一定关联，所以此方法的分类精度可能达不到实用。但相对实验目标，方法是有效的。

实验有两个过程，训练过程：从 $D1$ 中生成分类规则集 R；测试过程：以一定的计算方法，用 R 来对 $D2$ 中记录进行分类。

训练过程首先将训练集中的属性进行分类。对于连续属性，将简单地进行等间隔离散，属同一间隔的数值被划分为一类；对于字符属性或离散属性，取相同值的属性被划分为另一类。为了区分记录的类别，之后中属性的类将称为簇，记录的类别将称为类。接下来计算每个簇与类之间的相互支持的置信度，将符合阈值的规则作为分类规则。训练过程需要设定两个参数：充分置信度阈值和必要置信度阈值，另外要设置一整数 K，作为离散连续属性时的区间数。具体训练步骤如下：

（1）访问一次 $D1$，获取记录类集合和每类记录数；获取每个连续属性的最大值和最小值，由 K 计算出每个连续属性的离散区间大小；获取每个字符属性的所有不同取值。

（2）再访问一次 $D1$，统计每簇与每类的关联关系数据，得到一中间集合。

（3）依据中间集合中数据计算各簇与各类间的双向置信度，将满足阈值（≥）的规则记入集合 R，R 是训练结果，将在测试时作为分类计算的依据。

下面以一小训练集的训练过程来说明其中的细节。训练集 D 有 $A1$ 和 $A2$ 两条件属性和一个类别属性 C，其中 $A1$ 是连续的数值型，$A2$ 是离散的字符型。如表 4.1 所示。

表 4.1　　示 例 训 练 集 *D*

序号	$A1$	$A2$	C	序号	$A1$	$A2$	C
1	1855	优秀	$C1$	4	1722	良好	$C2$
2	2016	良好	$C1$	5	2588	良好	$C2$
3	1600	优秀	$C1$	6	2290	良好	$C1$

（1）训练时，设充分置信度阈值为 70%，必要置信度阈值为 90%，$K=3$。

第 1 次访问 D 得到类和簇的集合或划分，结果有：类别 C 的集合为 {0：

$C1$，1：$C2$}，其中$C1$记录数为4，$C2$记录数为2；$A1$的最大值2588，最小值1855，离散间隔=(2588−1855)/3=244，则$A1$的离散区间为{0：[1855，2099)，1：[2099，2343)，2：[2343，2588]}；A2取值集合为{0：优秀，1：良好}。

第2次访问D得到类和簇关联关系的统计数据，结果以属性簇为索引，如表4.2所示。其中的第一行表示：D中满足属性$A1$的0簇条件的记录有4条，其中属于$C1$的记录有3条，属于$C2$的记录有1条，后面的类似。

表4.2　　第2次访问D后得到簇与类关联关系的统计数据

序号	属性标识	簇标识	簇中记录数	其中$C1$记录数	其中$C2$记录数
1	$A1$	0	4	3	1
2	$A1$	1	1	1	0
3	$A1$	2	1	0	1
4	$A2$	0	2	2	0
5	$A2$	1	4	2	2

依据表4.2数据计算双向置信度，得到的分类规则按类索引，结果如表4.3所示，表4.3就是训练得到的规则集R。以表4.2的第一行为例，属性$A1$的0簇共有4条记录，其中有3条记录属于$C1$，而D中属于$C1$的记录有4条，那么“$A1$：0→$C1$”的充分置信度为3/4=75%，大于阈值70%；必要置信度3/4=75%，小于阈值90%，记为0。

表4.3　　D训练得到的分类规则表

序号	类标识	属性标识	簇标识	充分置信度	必要置信度
1	0 ($C1$)	$A1$	0	3/4=0.75	0
2	0	$A1$	1	1/1=1	0
3	0	$A2$	0	2/2=1	0
4	1 ($C2$)	$A1$	2	1/1=1	0
5	1	$A2$	1	0	2/2=1

(2) 测试时，分两种方法。

1) 方法1只考虑充分置信度。从D_2中取出一条记录t，离散化其属性。分别以t的属性匹配R中各类的规则。t属于c_j的得分按式(4.6)来计算，即匹配上的规则，计算其充分置信度平方和。将t判定为属于得分最高的类。

$$\text{SCORE1}(c_j,t)=\sum_{\text{匹配}}\text{Confidence}^2 \tag{4.6}$$

2) 方法2考虑充分置信度和必要置信度。从D_2中取出一条记录t，离散

化其属性。分别以 t 的属性匹配 R 中各类的规则。T 属于 c_j 的得分按式(4.7)来计算，即匹配上的规则，计算其充分置信度平方和；匹配不上的规则，计算其必要置信度的平方和；以前者与后者的差作为 t 支持 c_j 的得分。将 t 判定为属于得分最高的类。

$$\text{SCORE2}(c_j,t)=\sum_{\text{匹配}}\text{Confidence}^2-\sum_{\text{不匹配}}N\_\text{Confidence}^2 \tag{4.7}$$

以记录 $X=\{2300,\text{良好}\}$ 为例进行分类测试。首先识别 X 中属性的簇，结果为 $X=\{A1:1,A2:1\}$；扫描表 4.3，X 匹配其中的 2，5 两条规则，那么：

按方法 1，$\text{SCORE1}(C1,X)=1^2=1$；$\text{SCORE1}(C2,X)=0$；$X$ 被判定为 $C1$ 类；

按方法 2，$\text{SCORE2}(C1,X)=1^2=1$；$\text{SCORE2}(C2,X)=-1^2=-1$；$X$ 被判定为 $C1$ 类。

应该说明的是，此训练与分类方法有很多缺陷，不能成为实用的算法，但对于实验目标是有效的。

4.4.3 在 UCI 分类集上的测试

本节实验数据集是来自 UCI 机器学习库[116]的 4 个分类集：Mushroom，Wine，Zoo 和 Breast。以每个数据集作为训练集，再以自身作测试集。设置充分置信度阈值为 50%，必要置信度阈值为 80%，测试结果如表 4.4 所示。

表 4.4　　UCI 分类集上的测试结果

序号	训练集	测试集	数据集规模	分类精度	
				方法 1	方法 2
1	Mushroom	Mushroom	8124	0.582	0.597
2	Wine	Wine	178	0.567	0.567
3	Zoo	Zoo	101	0.703	0.851
4	Breast	Breast	699	0.961	0.973

由表 4.4，在考虑了必要置信度后，分类精度普遍得到了提高。但 Wine 数据集的测试精度不变，研究其生成的规则集 R，发现其中根本没有满足必要置信度阈值的规则。再分析 Wine 数据集，它有 13 个属性和 178 条记录，其中 12 个是连续值的条件属性，属性值分布较均匀。另一个属性是类别，类别有“1”、“2”、“3”3 类，记录数分别是 59、71、48。在实验时属性的离散化采用的是等间隔离散，间隔数是 10，由式（4.5）可以看出为什么没有采集到满足必要置信度阈值（$\geqslant 80\%$）的规则。这组实验数据表明，如果能采集到合适的必要规则置信度，并让它们在分类时起到适当的作用，分类精度会提高。

4.4.4 双向置信对不平衡数据集分类的测试

针对不平衡数据集的分类问题是一类重要的分类问题，在网络入侵检测、信用卡欺诈识别及疾病诊断等领域有实际应用。这类问题的数据分析有3个层面：

(1) 在不平衡数据集中识别出稀有类记录。例如，在网络访问数据集中识别出攻击记录，这类记录一般占很小的比例（如低于10%），称为稀有类。这一层面的问题一般称稀有数据挖掘问题，其挖掘方法一般是基于距离或密度的，如本书第3章所述。

(2) 已知某记录是稀有类记录，识别出其属于稀有类中的哪一类。例如，在网络访问数据集KDDCUP99中，正常访问记录是大类，攻击记录是稀有类，攻击记录又分为22个类。已经知道某记录是攻击记录，识别它是哪类攻击是第二个层面问题。

(3) 在不平衡数据集中识别出大类和稀有类，并识别出稀有类的类别。可以看出这3个层次是逐次深入的，第3.4.4节的方法和测试是针对第三个层面的。

当然稀有数据挖掘算法能有效，说明不平衡数据集中稀有记录的属性确有异常的表现。仔细分析KDDCUP99数据集，可以发现其中的攻击记录属性具有特异性，并且不同类的攻击记录其特异属性有明显差距。基于这一事实，并考虑到不平衡数据集中起决定作用的是稀有属性（支持度小于等于指定阈值的属性）；一记录被判定不是某个攻击类，那么它就是正常类，因此对第3.4.2节的实验方法进行了以下修改：

(1) 设置一支持度阈值，如10%，在生成规则集时只采集前件支持度≤10%，并且满足两置信度要求的规则，不采集正常类的规则。

(2) 在测试时，t仍被判定属于得分最高的类。由于不采集正常类的规则，只要最高得分>0，就判定是攻击类，当所有攻击类得分均小于等于0时，判定t为正常类。

1. 识别精度测试

KDDCUP99是网络访问记录数据集，每行数据记录了一次网络访问的属性及访问类别。访问类别反映了本次访问的性质，属于典型的不平衡数据集，其中大部分记录属正常访问类（normal），称为大类，例如>90%。另外的类别均属于攻击类，称为小类。我们从KDDCUP99中选择了corrected数据集，corrected数据集共有65536条记录。从中选择了正常类记录和8类攻击记录组成4个数据子集，Sub1、Sub2、Sub3、Sub4，其记录构成如表4.5所示。各子集中攻击记录的比例均<10%，Sub1和Sub4中攻击类记录是均匀的，Sub2和Sub3中攻击类记录是非均匀的。

表 4.5　　4 个子集的记录构成

子集名	其中含各类记录数									记录总数
	apache2	back	neptune	ipsweep	portsweep	saint	smurf	Guess _ password	normal	
Sub1	10	10	10	10	10	10	10	10	1000	1080
Sub2	20	17	9	10	15	5	8	4	1500	1588
Sub3	30	18	4	15	28	8	5	11	2000	2119
Sub4	30	30	30	30	30	30	30	30	6308	6548

设置支持度阈值为 10%，充分置信度阈值为 50%，必要置信度阈值为 80%。分别以各子集为训练集和测试集，进行实验，结果如表 4.6 所示。表中错判记录数由 3 列合计得到，“攻-攻”表示其中攻击记录被错判为另一类攻击的记录数，“攻-正”表示其中攻击记录被错判为正常类的记录数，“正-攻”表示其中正常记录被错判为攻击类的记录数。

表 4.6　　KDDCUP99 数据集上的测试结果

序号	训练集	测试集	测试集规模（正常记录+攻击记录）	方法 1 错判记录数及精度				方法 2 错判记录数及精度			
				攻-攻	攻-正	正-攻	精度	攻-攻	攻-正	正-攻	精度
1	Sub1	Sub1	1080（1000+80）	0	0	38	0.965	0	0	0	1
2	Sub2			5	0	36	0.962	3	0	12	0.986
3	Sub3			8	0	33	0.962	0	0	6	0.994
4	Sub4			0	0	18	0.983	0	1	5	0.996
5	Sub1	Sub2	1588（1500+88）	3	0	66	0.957	0	1	0	0.999
6	Sub2			0	0	44	0.972	0	0	4	0.997
7	Sub3			5	0	45	0.969	0	0	1	0.999
8	Sub4			0	0	30	0.981	0	1	1	0.998
9	Sub1	Sub3	2119（2000+119）	9	0	86	0.955	0	6	0	0.997
10	Sub2			8	0	69	0.964	6	3	14	0.989
11	Sub3			2	0	65	0.968	0	1	6	0.997
12	Sub4			2	0	43	0.979	0	4	5	0.996
13	Sub1	Sub4	6548（6308+240）	21	1	175	0.970	1	16	0	0.997
14	Sub2			25	0	140	0.975	17	11	33	0.990
15	Sub3			25	0	132	0.976	2	8	17	0.996
16	Sub4			3	0	80	0.987	1	8	19	0.996

由表 4.6，所有测试结果中方法 2 的精度均高于方法 1 的精度，方法 2 的精度非常高；从 4 项一组的实验中可以看出训练集的规模、训练集中小类记录

的均衡程度对测试结果影响较小；由于不采集大类的规则，实验中生成的规则集 R 规模较小。

表 4.6 中显示的是整个数据集的分类精度，在不平衡数据集中，小类是被关注的对象，其被识别的精度更能反映算法的性能判别。表 4.7 反映的是以 Sub1 为训练集时，Sub2 测试结果中每个攻击类的正确率 P、召回率 R 和 $F1$ 值，其中对比了方法 1 与方法 2 的值。P、R 和 $F1$ 计算公式如下：

$$P=\frac{\text{正确分为某类的记录数}}{\text{测试集中分为该类的记录数}} \tag{4.8}$$

$$R=\frac{\text{正确分为某类的记录数}}{\text{测试集中属于该类的记录数}} \tag{4.9}$$

$$F1=\frac{2\times P\times R}{P+R} \tag{4.10}$$

表 4.7　　Sub2 测试中各攻击类的 P、R 与 $F1$ 值对比

类别	方法 1			方法 2		
	P	R	$F1$	P	R	$F1$
apache2	1	1	1	1	0.9	0.947
back	0.548	1	0.708	1	1	1
neptune	0.643	1	0.783	1	1	1
ipsweep	0.833	1	0.909	1	1	1
portsweep	0.316	0.8	0.453	1	0.933	0.966
saint	0.385	1	0.556	1	1	1
smurf	0.5	1	0.667	1	1	1
Guess _ password	0.364	1	0.534	1	1	1

由表 4.7，在适当利用必要规则置信度后，P、R 和 $F1$ 的值均得到了显著提高。

2. ROC 曲线分析

本实验中，方法 1 与方法 2 判定一记录类别时，均依据量化的得分值。在 4 个数据子集 Sub1、Sub2、Sub3、Sub4 中，包含 8 个攻击类和一个正常类。如果以一个攻击类为正类，其他攻击类和正常类为负类。那么，据量化的得分绘制 ROC 曲线，可以有效表征两方法的分类效果，从而说明必要置信对分类的贡献。仍以 Sub1 为训练集，以 Sub2 测试结果中 portsweep 攻击类的方法 1 与方法 2 得分为数据，绘制 ROC 曲线，如图 4.1 所示，其曲线下面积的对比如表 4.8 所示。

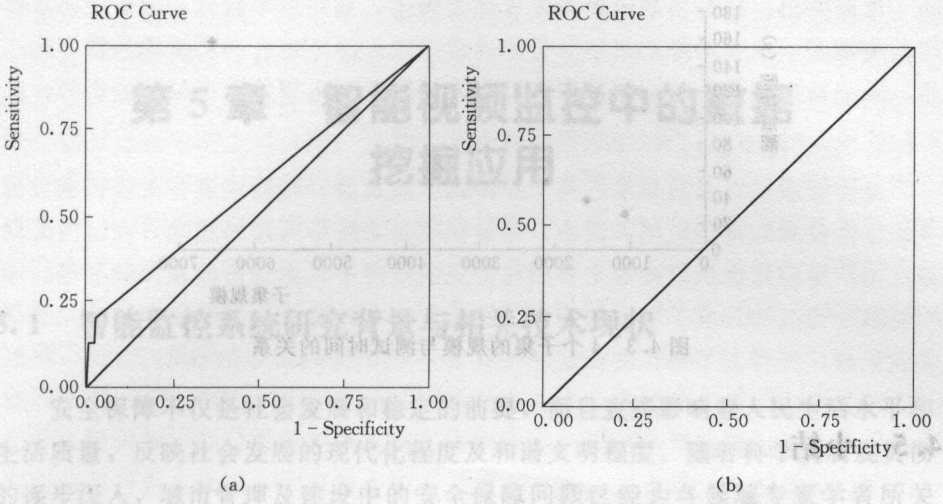


图 4.1 portsweep 类在测试中得分的 ROC 图

(a) 方法 1；(b) 方法 2

表 4.8 portsweep 类测试得分 ROC 曲线下面积

	方法 1 ROC 曲线下面积	方法 2 ROC 曲线下面积
portsweep 类	0.589	0.997

由图 4.1 和表 4.8 可知，必要规则置信度对 portsweep 类的识别有显著贡献。

3. 时间效率分析

本实验方法的训练过程只需访问两次训练集，其时间复杂度为 $O(N)$，N 是训练集规模。测试过程，方法 1 与方法 2 均需访问一次测试集，对测试集中的每个记录，需与规则集的每个规则匹配，其时间复杂度是为$O(KN)$，其中 K 是规则集中规则的数目，N 是测试集的规模。在实验中也验证了这一点，不同规模数据集的训练时间、对应方法 2 的测试时间如图 4.2 和图 4.3 所示。

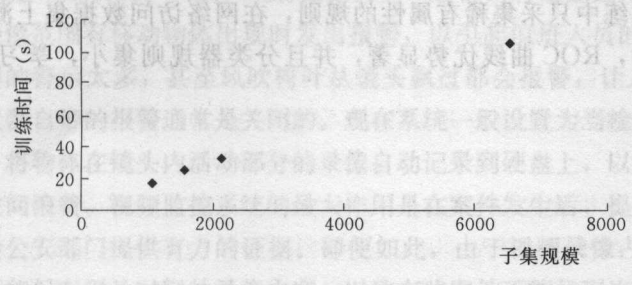


图 4.2 4 个子集的规模与训练时间的关系

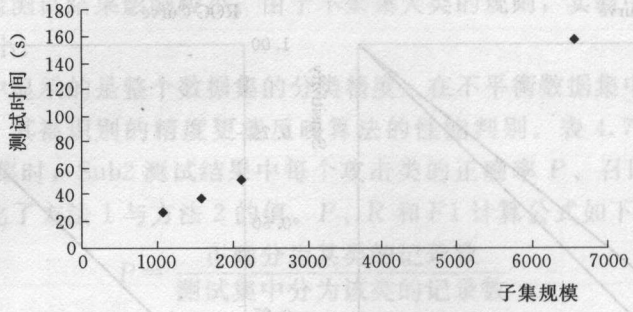


图 4.3 4个子集的规模与测试时间的关系

4.5 小结

分类一般有两个步骤：

（1）从训练集中找出规律，这一步也称为有监督的机器学习。

（2）依据找出的规律，判断测试集中记录的类别。

C4.5算法、CBA算法等基于规则的分类算法从训练集中找出的是形如“$A \rightarrow C$”的规则，其中，A表示全部或部分属性的一些取值组成的集合，C表示某个类标号。这类算法基于的思想是：“如果规则‘$A \rightarrow C$’的置信度达到一定阈值，则认为A的出现增加了结果为C的概率或可能性。”我们认为：“如果规则‘$C \rightarrow A$’的置信度达到一定的阈值，则可以认为A不出现减少了结果为C的概率或可能性。在分类算法中如果同时考虑‘$C \rightarrow A$’的作用，会提高分类精度。”

设计了两类分类算法，方法1只考虑“$A \rightarrow C$”的影响，方法2考虑“$A \rightarrow C$”和“$C \rightarrow A$”的影响。分别在UCI机器学习库的4个分类集Mushroom，Wine，Zoo、Breast上进行了实验测试。结果表明：如果能采集到合适的“$C \rightarrow A$”规则，并让其在分类中起作用，可以有效提高分类精度。

特别值得说明的一点，考虑到不平衡数据集中起决定作用的是支持度小的稀有属性，训练中只采集稀有属性的规则，在网络访问数据集上测试，方法2分类精度很高，ROC曲线优势显著，并且分类器规则集小，学习时间与测试时间成线性。

第5章　智能视频监控中的数据挖掘应用

5.1　智能监控系统研究背景与相关技术现状

安全保障不仅是社会发展和稳定的前提，而且直接影响着人民生活水平和生活质量，反映社会发展的现代化程度及和谐文明程度。随着科学发展观贯彻的逐步深入，城市管理及建设中的安全保障问题已经为各领域专家学者所关注，并且越来越多地引起政府管理部门的高度重视。安全保障的内涵很广泛，其中安防系统是安全的重要保障。我国城市街道、广场、居民生活小区、学校等区域均配备了相应的安防系统，并且还在逐步加大投入。一方面，城市区域管理、保安人员配备及人员素质正在逐步完善及提高；另一方面，安装的安全监测、预防设备也越来越先进，比如目前的远红外周边报警系统、视频监控系统、电子巡更系统和门禁对讲系统等。

视频监控系统以其直观、方便、信息内容详实，被广泛应用于生产管理、保安等场合。视频监控系统的一般过程是：在一些重要的场所安放一个或若干个摄像机拍摄监控现场，然后将视频信号通过一定的传输网络，传到指定的监控中心，再存储到存储介质上，同时还可以根据不同需要和途径在现场安装其他的探测装置作为监控系统的辅助设备。我国目前大部分城市街道、居民生活小区及学校的重要部位都安装了视频监控设备，以实时了解动态情况，在一定程度上起到防范和管理的作用。但现有的视频监控系统仍存在较多问题。首先，监控中心每天24h有人值班，但是，要求值班人员时刻紧盯着几十个屏幕，随时发现可疑情况是不现实的。其次，现有摄像头上一般都具有诸如红外报警等设施，可以设置当摄像范围有移动物体出现时发出报警，以引起值班人员的注意。但是，这种报警机制的警示太多，甚至风吹树叶从镜头飘过都会报警，让人感到不堪其扰。因而，仪器自带的报警通常是关闭的。现在系统一般设置为当检测到镜头下有移动物体时，将物体在镜头内活动部分的录像自动记录到硬盘上，以避免24h记录录像的硬盘空间浪费。视频监控系统的最大作用是在案件发生后，提供当时现场的录像资料，给公安部门提供有力的证据。即便如此，由于视频录像占用空间极大，监控中心也不能保存很长时间的录像内容，以致有些案件不能找到当时的录像。

可以看出，现有视频监控系统以硬件为主，利用效率较低，而系统深层智能的关键在于视频录像的自动分析和异常发现。由于现有的系统均为数字化产品，其记录在硬盘上的录像数据可以作为计算机深入智能数据分析的资源。

5.1.1　安防技术现状

受需求驱动，安防技术近几十年得到了大发展，指纹识别、人脸识别、数字水印等技术均得到应用。其中，指纹识别是安防中的重要技术，在国内外均有应用。目前指纹识别技术主要用在公民的证件中，如身份证和保险卡等。康柏电脑公司成功开发了电脑用指纹识别装置，有效地防止了电脑内的信息被盗取或更改，并省去了操作者设定并记住那一长串密码口令的麻烦[118]。一种通过三维扫描来识别人脸的技术在英国机场等地用来识别被通缉的罪犯[118]。指纹识别和人脸识别技术主要基于高精度、近距离的影像或视频。英国南安普敦大学研究人员指出，除非经过刻意训练，人的走路姿态一般难以进行伪装。因此，步态有可能起到与面容类似的识别作用，以此可帮助增强技防和破案能力。英国科学家正在研制一种通过分析走路姿态以识别嫌疑犯身份的安防系统。步态分析的好处是证据容易收集，使用低分辨率摄像机便可从远距离、任何角度拍摄下可供分析的步态图像[118]。此外，利用数字水印防止伪造证件，利用网络监控住宅及公共重要场所等应用也较广泛。

5.1.2　相关技术研究现状和发展趋势

视频摄像的智能研究涉及的关键技术有：图像特征提取与视频识别技术、数据挖掘技术、模糊数据挖掘及在视频识别中的应用等。相关技术已经过相当长时间的研究与发展，有一些智能监测报警系统成功应用，许多理论、方法与应用可以借鉴，以下分别阐述。

1. 图像特征提取与视频识别技术方面

图像特征提取是图像识别、视频识别、计算机视觉的基础步骤，国内外最常用的是边缘检测方法，经典的边缘检测算法在MATLAB等软件中均已实现。针对经典算法的缺陷，即只能针对局部给出模糊边缘信息和对噪声及纹理敏感，一些学者提出了改进方法。如张浩等利用多个边缘算法的输出进行贝叶斯推理，判断每个像素是否属于边缘，从而实现复杂场景的边缘检测[119]；罗敏等提出了一种基于径向小波变换的图像特征提取算法，该方法使用径向对称函数来获得图像的边缘信息，算法容易实现，能够用于基于内容的图像检索[120]。

在很多计算机视觉应用中，一个基础而关键的任务是从视频序列中确定运动目标，其中对于固定摄像机的监控视频运动目标的检测，最常用的方法是减背景技术。其思想是将视频帧与一个背景模型做比较，其中区别较大的像素区

域被认为是运动目标。但由于构建背景模型需要考虑光照变化等很多因素，因此开发一个好的减背景算法面临很多挑战。代科学等在文献［121］中对利用减背景技术实现运动目标检测的过程、各种典型背景建模算法的原理和优缺点做了较为详细的阐述和归纳，总结了各种减背景算法的总体特点，指出了减背景技术的未来研究重点和发展方向。从视频中识别人的姿势和动作有很多潜在的应用，如自动监控、疾病诊断、运动分析和人机交互等。姿势识别的困难在于图片的混杂性和人体多关节结构。姿势的识别可以由识别人的头、四肢等组成部分来进行。文献［122］中应用了一种基于马尔可夫链的数据驱动方法。由肢体的方向形成3D姿势估计模型，再到事先存放的姿势候选集中进行马尔可夫匹配搜索，形成最终的姿势模型。J. K. Aggarwal 等在文献［123］中描述了一些姿势识别高级处理技术的方法与现状，如：人体建模、理解人的行为的细节、人类运动识别的方法、在领域中的高层模式识别等。在视频识别中，车辆识别的研究最多[124]，其理论与方法在智能交通中应用广泛。

2. 智能监测报警应用系统方面

在监测应用系统方面，Ismail Haritaoglu 等开发了一个在室外录像中识别多人及他们的动作的系统——W4[126]。系统利用形状分析、识别人的构成部分来建立人的外观模型，将有多人的前景进行分段，在有多人互相遮挡的录像中跟踪多人的交互。系统还能将人携带的物品与人分离，识别对物体的动作。Oberli C 等开发了一个专家系统来从对做过心脏病手术的病人在线监测数据中发现异常并报警[127]。系统针对的问题是原监测数据没有被充分利用，原有仪器的错误报警使得人们对警示不再在意。系统采用模糊判断逻辑来适应不完全和有噪声的数据。与监测仪器自带的报警系统相比，原75%的仪器报警是误报，此专家系统的误报率低于1%；对异常情况的感知仪器是79%，专家系统是92%；对积极情况的预测能力仪器是31%，专家系统是97%。可见此专家系统有效地提高了原检测仪器的使用能力和效率。英国 PC PRO 网站2008年6月25日最新的一则新闻显示，英国朴次茅斯科学大学正在进行一项3年的CCTV智能化项目[128]。项目致力于让CCTV的摄像头“听”出暴力声音，如窗子破碎的声音、人尖叫与咒骂声音等，并让镜头转向声音的方向。开发者欲从现场的声音波形中进行模糊数据分析，以识别异常。

索尼公司新近推出了 IMZ－RS400 系列智能监控软件[129]。该软件与索尼配套的视频网络联合使用，主要实现了6种基本报警机制。这6种报警机制是：

（1）穿过—用户可设置一条虚拟的边线，当有物体穿过边线时报警，可用在如公路上的安全栏杆边的监控，当有人翻越时报警。

（2）出现—设置一个虚拟区域，当有物体出现其中时报警，可用在禁人、

物进入区域的监控。

(3) 消失—当虚拟区域中有目标物体消失时报警，可用于重要物品防盗监控。

(4) 容量—当虚拟区域中目标物体的数量超过一个指定数值时报警，可用在如博物馆、电影院等公共场所人数控制报警。

(5) 存在—当目标在虚拟区域中存在的时间超过预定值时报警，如用在禁止停车区域的监控。

(6) 离开—当无人看管目标物体离开虚拟区域超过预定时间时报警。

可见，IMZ-RS400系列智能监控软件抽象了现实中一些场合的异常模式，针对这些模式设计报警机制。系统中只识别人、物的进入、消失、移动方向、速度等，没有识别动作、频率、轨迹等行为。

中国科学院深圳先进技术研究院智能仿生研究中心研究开发的智能监控功能较IMZ-RS400系统软件又有了进步[130]。在其识别特异行为的功能中，可以看到数据挖掘应用的效果。北京智安邦科技有限公司推出的ZANB智能视频监控系列产品已实现在真实的环境中警戒区入侵检测、警戒线穿越检测、物品被盗或移动检测、早期火灾检测、人员异常聚集、倒地检测等功能[132]。该公司还为不同的行业提供完整的解决方案，如：部队仓库及重要单位智能安防系统、火车驾驶员疲劳检测系统、楼宇智能视频监控报警系统、机场安防消防智能视频监控报警系统、无人值守设施智能视频监控报警系统、周界（围栏、围墙）智能视频监控系统、石油石化企业智能视频监控系统、仓库安防消防智能视频监控系统、博物馆和文物保护单位安防消防智能视频监控系统、监狱智能视频监控系统等。另外，如Panasonic[133]，Vidient[134]，VistaScape[135]也均有商业的智能监控系统推出。

但仔细研究这些产品的功能就会发现，现有的视觉智能主要是对已知目标或指定行为进行识别，应用也限定在一些特定的场合，目前智能交通的应用比较广泛[131]。对于人的行为精确识别与理解还有待研究与开发。一些国际的学术期刊与学术会议也对相关研究进行了集中展示，以方便学习和交流[136]。但由于现实世界的复杂性和人的行为的动态性，相关研究还面临许多挑战，如：

(1) 现实的背景是复杂多变的，目前的研究成果要走出实验室还有些距离。

(2) 由于多个人之间可能会有遮蔽、有的人可能会有部分部位出现在镜头之外、有些人可能穿着宽松的衣服等，识别人的各种行为技术还不成熟。

(3) 视频识别计算量很大，要达到实时和预测还需要深入研究。

图像特征具有模糊性，人对图像的识别也具有模糊性，模糊理论在图像处理中应用较有实效。数据挖掘、模糊挖掘理论与技术已得到了大发展，相关算

法效率已达到实用。数据挖掘在视频识别中的应用研究一般称为视频挖掘。视频挖掘技术的目标是实现视频图像的低级特征向高级语义信息的转换，从大量视频数据中自动提取隐含的、有用的、可以理解的模式或知识，为人们提供问题求解层次的智能或智力支持[138]。目前美国和日本均有小组专门进行视频挖掘的研究[138]。利用模糊数据挖掘技术研究运动目标的特征提取、运动目标行为识别、趋向判断和异常检测，将有力促进数据挖掘、模糊理论与图像处理技术的融合。

5.2 一种智能监控系统构架

5.2.1 一种智能监控系统构架描述

目前的技术直接识别人的各种行为还不可能。本小节提出的构架目标是识别录像中的异常情况，主要识别指定的行为模式，通过学习逐步积累行为模式，使系统功能不断完善。其构架如图 5.1 所示。

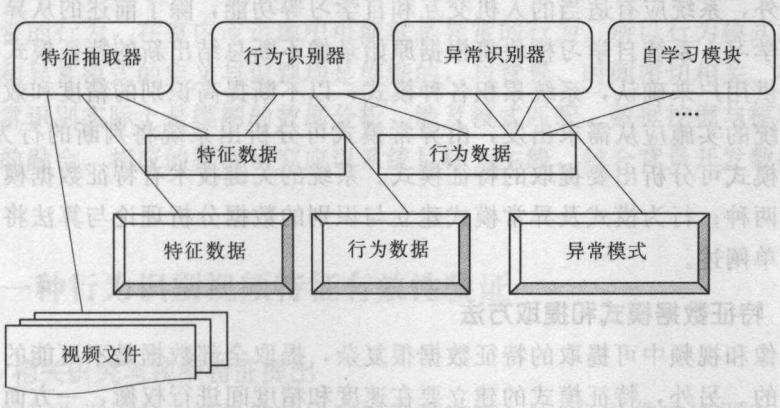


图 5.1 一种智能监控系统的构架图

系统的数据由 3 层构成，即原始录像的视频文件、模式和实时数据。模式主要有特征模式、行为模式和异常模式；实时数据有特征数据与行为数据。

特征模式反映要从录像中提取的数据特征。人或物体的运动特征是提取的重点，如出现、移动物体个数、外形、运行方向、运动轨迹、人的四肢摆动速度与方向等。系统根据欲提取的特征建立特征模式，特征提取模块根据特征模式从录像中提取实时的特征数据。

行为模式反映移动物体的行为，如越线、消失、快速奔跑、暴力等。行为识别模块主要根据系统定义的行为模式，识别特征数据的行为，生成实时的行为数据。

异常模式反映行为数据中的异常。异常与时间、地点有关，开发过程中首先应根据系统背景对异常行为进行抽象。例如指定时间与指定地点的出现。如财务室、档案室、票据室、网络中心机房等场所，在午夜到凌晨时间段出现人物均可视为异常；指定时间与指定地点的越线。如设在窗口的镜头，可在窗前设置一条虚拟的线，当有人穿过这条线可视为异常；暴力行为等。

异常识别模块从3个方面定义与识别异常：

(1) 根据背景抽象出的异常行为定义为异常模式，符合异常模式的行为数据被识别为异常行为。

(2) 使用数据挖掘的异常挖掘算法来识别异常情况，数据挖掘中的异常挖掘可以从大量记录中识别每个记录的异常程度，将异常程度高的行为识别为异常。

(3) 系统通过积累不断进行自学习，如从过去已知是异常情况的片断中学习，或者通过人机交互由用户将过去片断中的行为模式定义为异常或正常。当然，对怀有异常目的，但没有行为表现的情况，系统是无法识别的。

此外，系统应有适当的人机交互和自学习等功能，除了前述的从异常片断中的自学习，系统自学习模块将根据原始录像不断总结出新的特征模式、行为模式，供用户来确认，系统累积各种模式，以不断提高识别的精度和效力。

系统的实施应从需求出发，由异常模式可分析出系统将判断的行为模式，由行为模式可分析出要提取的特征模式。系统的关键技术有特征数据模式和提取方法两种；行为模式及异常模式建立与识别的数据分析理论与算法将在以下进行简单阐述。

5.2.2　特征数据模式和提取方法

图像和视频中可提取的特征数据很复杂，提取全部数据是不可能的，也是不必要的。另外，特征模式的建立要在速度和精度间进行权衡，一方面，提取尽量多的数据可以少丢失录像中的信息；另一方面，系统的输入录像是多路的，要保证实时性和效率，就应尽量丢掉无关的信息。

研究提取视频特征初步可归为两大类，一类是为识别人或物体位置、方向、速度等而提取的特征。另一类是为识别人的动作、行为而提取的特征，这类特征模式定义及数据提取较复杂。特征模式建立及数据提取应分几个层次进行：

(1) 当有移动物体进入视频范围时，提取其位置、运行方向、速度、轨迹和外形等特征。

(2) 当有移动物体进入视频范围时，提取其外形及活动部分特征，以判断其是人，还是其他物体。

(3) 当单个的人进入视频时，提取其头部、躯干、四肢的位置、方向及速

度等特征，以识别人的简单行为。

（4）当多人进入视频时，提取他们未被遮蔽部分的特征数据，以识别个人行为和互行为。

因为人有各种形态、进入视频时可能从不同角度、人可能骑着自行车或携带物品、多个人可能互相遮蔽等各种情况，定义特征模式及提取数据非常复杂。模糊识别与数据挖掘技术会大有用武之地。

5.2.3 行为模式及异常模式建立与识别的数据分析理论、算法

系统中的行为模式识别有两个方面：

（1）根据系统现存的行为模式识别特征数据的行为。

（2）当特征数据不能与现有行为模式匹配时，总结生成新的行为模式。

行为模式是由特征数据定义的，但除了越线、出现、消失等简单行为，很多行为的特征数据是模糊的。例如，“跑”和“走”是两个行为，人区别“跑”和“走”的直观特征是速度。但移动速度达到多少就是“跑”而不是“走”呢？这个问题很难有一个硬的界线。再如，对同一速度的运动，在室外的广场上是正常的行为，在室内或走廊中可能就是异常的行为。所以行为模式的定义与识别和异常模式的定义与识别均是模糊的，与背景、时间密切相关。模式识别和异常识别实际上是模糊的数据分析，涉及模式匹配、数据挖掘与模糊理论及方法的融合。相应的算法研究应以系统目标为主线，结合背景开发数据分析算法。

5.3 一种行为识别视频特征有效性验证

5.3.1 相关研究与视频特征提出

人的行为识别是计算机视觉研究的重要内容，从2D录像中识别出人的行为是智能视频监控的基础任务，近年来有大量的研究成果[139]。综合已有的研究，从2D录像中识别人的行为需两个关键步骤：

（1）从2D录像中提取适当的特征数据。

（2）采用适当方法进行识别。

从视频中提取的特征数据是后续识别的基础，也是各研究互相区别的主要特征。根据数据特征的不同，可将已有的行为分析方法分为3类：

（1）基于姿势变化来识别行为的方法。该类方法先从视频中提取出组成人体的各部分，如：头、四肢、躯干等，识别人的姿势序列，再识别人的行为[139]。但由于遮挡、光线、视角等因素变化，从2D视频中准确提取人的组成部分相当困难，姿势识别的难度在一定程度上超过了行为识别。多摄像头数

据融合是目前解决遮蔽问题的主要思路，但其计算复杂，实时处理困难[139]。

(2) 基于视频前景中人体区域时空特征的模式识别方法。这类方法提取视频中人体的区域，利用区域、轮廓的特征来描述和识别行为。如：Catherine等主要利用人物轮廓的2阶矩来识别人的行为[141]。胡芝兰等利用视频中前景块的运动方向特征来检测公共场所异常行为，对单人及多人场景均可识别，计算复杂度小，但其方法不识别人群中某个人的具体行为[139]。Lena Gorelick等将每帧的运动轮廓在时间轴上连接起来，形成三维体。利用三维体的外形、方向等特征来识别运动[142]。总体上讲，这类方法避免了精确提取人体部位的困难，时间效率和鲁棒性较好。

(3) 基于视频帧时空特征的模式识别方法。Bobick等用一种静态图片向量来表示运动模型，向量中的每一个点由包含此点运动特征的函数构成[143]。Ivan Laptev等利用兴趣点（interesting points）的特征来识别人的行为，识别算法只利用了视频中的几个兴趣点的特征，但兴趣点是在全部视频序列图像中梯度变化大，并且其周围点梯度变化也大的点，其搜索算法较复杂[144]。此类方法利用视频的全部特征，前景和背景，一般来说比第2类方法计算量大。

不同的应用背景对识别特征与识别算法的要求不同，但有一些要求是共同的，如：特征易于提取；特征视角不变；识别精度高；识别效率高；适应复杂变化的环境；算法具有鲁棒性等。虽然有很多行为特征和识别算法被开发出来，但由于视频与环境的复杂性，准确识别人的行为的特征与算法还需要更深入和广泛的探索研究。

仔细观察各类运动的2D视频会发现，不同行为在一定程度上表现为人身体不同部位的伸与缩，不同部位的伸与缩在横向上表现为人不同高度区间的伸与缩。同时，在纵向上表现为不同纵向区间的伸与缩。例如：走路时，手臂直臂前后摆动，腿和脚基本伸直前后摆动，整个人形宽度伸缩大的部位在腰以下，以及脚的部位；跑步时，手臂端在腰间前后摆动，整个身体宽度伸缩大的部位在腰间，以及膝和脚的部位。不同的运动，在纵向上也可以观察到不同的伸缩变化。是否可以利用人整体横向和纵向不同区间的宽度和高度伸缩来表征不同的运动行为呢？将以实验验证该特征对行为识别的有效性。

5.3.2 实验设计

1. 实验素材与视频处理软件

实验对象是固定摄像头下单人运动视频录像，选用Christian Schuldt等制作的行为视频作素材[146]。该视频数据库称为KTH行为数据库，包括6种行为，分别是走、慢跑、跑、拳击、挥手和拍手。每类行为视频分别由25人，在4类不同的背景下录制。第1类是静态稳定的背景；第2类摄像头是固定的，但是镜头不断地拉近与拉远，录像范围变大变小；第3类录像中人穿了不

同的衣服；第4类整个画面的光照不断变化。视频25帧/s，每帧120×160像素。

openCV是由Intel公司资助的开源计算机视觉函数库。它由一系列C函数和少量C++类构成，实现了图像处理和计算机视觉方面的很多通用算法。openCV的功能包括处理图像、视频和各种动态数据结构的操作，也包括相关的分析与识别算法，如对光流、运动分割、跟踪的分析以及HMM模型等。视频前期处理采用openCV的基础功能实现。

2. 获取前景序列段

从视频文件中提取出运动前景序列是第一步工作。首先，对视频帧进行高斯平滑，以清除每帧中的细小噪声，而保留每帧的灰度分布特征。之后，对不同类型的视频采用不同方法提取前景。减背景技术是固定摄像头录像中取出运动前景的最直观的技术[121]。在实验视频中，“走、慢跑、跑”的视频一般是人从画面一端开始运动，从画面的另一端出去。实验中采用帧差累积法动态生成与更新背景，再采用减背景技术获取运动的前景帧序列；实验视频中的“拳击、挥手、拍手”3类视频是人站在画面中进行的，主要是手臂运动，如果采用帧差累积来生成与更新背景，减背景后只能获取手臂的运动，躯干部分没有了。注意到实验视频的背景以淡灰色为主，且大部分进行表演的人穿着深色衣服，所以对后3类视频简单地采用了固定颜色域值对视频帧进行分割。最后，对得到的前景序列进行形态学腐蚀和膨胀以平滑前景的边缘和填补小的空洞，这样获得了待处理的前景帧序列。如图5.2所示显示了实验中获取的“走”和“挥手”两类视频片段和前景序列。可以看出人的肤色与背景很接近，获得的前景帧中人脸与手等裸露的部分缺失了，但基本的人形是完整的。

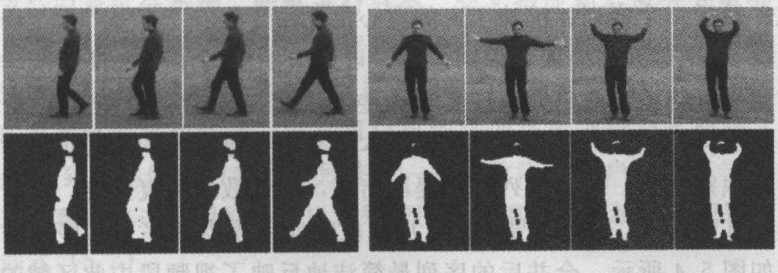


图5.2 “走”和“挥手”的视频片段及实验中获取的前景序列

“走、慢跑、跑”的视频，人从画面一端进入，从画面的另一端出去，有自然的分段，每段一般为15～35帧；“拳击、挥手、拍手”3类视频没有自然的分段，实验中按每30帧一段进行了划分。由于实验中获取的前景技术较粗糙，一些视频无法获取较完整的人形，不能采用，这样最终获得的实验视频前景片段为640段，其中：走，36段；跑，46段；慢跑，44段；挥手，229段；

拍手，131段；拳击，154段。

3. 提取特征数据

对获取的每一前景帧，实验中提取人形在横和纵两个方向上不同区域的宽度及宽度内部的空档宽度（纵向即为高度）。首先找出人形的左右和上下边界线，形成人形矩形；将矩形在横竖两个方向分别均匀划分为若干区间；对每个横向区间，找出最宽的一行像素，取其宽度，并取得这一行中内部空档（背景区域）的累计宽度，为适应远景和近景的变化，将得到的宽度与当前矩形的高度相比，得到相对人高度的值，下面指的宽度均是指此相对宽度；纵向数据按相似的方法获取；为了反映人整体运动的方向与速度，找出每帧人形的质心坐标 x 和 y。以横竖两个方向分别划分为5个区间为例，其示意图如图5.3所示。因为每个区间有最大宽度和最大宽度中的空档宽度两个数据，再加上质心坐标，每一帧获取的特征数据个数为：5×2+5×2+2=22。

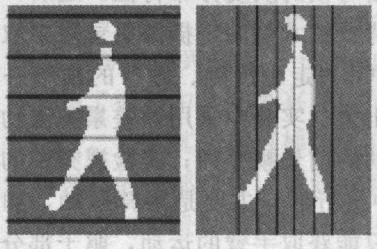


图5.3　将一前景帧分成5行5列示意图

仍以每帧横竖两个方向分别划分为5个区间为例，一段视频将产生22个序列。计算这22个序列的帧间差，即从序列的第2项开始，用每个值减去序列中的前一个值，得到的仍是22个序列值，只不过序列的长度少了1。这22个序列中的前20个序列中，正值表示此区域当前帧较上一帧变宽了，数据为负则表示当前帧较上一帧变窄了。因此序列反映了视频中人运动时横向、纵向各区域宽窄变化，同时内部空档的宽窄变化；最后2个序列，反映了人的质心位置的变化。

以这22个序列来表征和识别运动，可以采用隐马尔可夫模型或动态时间规划方法来进行，目前研究也证实了这些方法对视频识别的有效性[140]。由于时序匹配的存储量大，识别算法复杂，本节采用了更简单的模式识别方法。为适应模式识别方法，需将序列特征提取出来，降低数据的复杂度。对于每个序列，忽略其中的0，进行同号合并，形成反映变化拐点的序列，序列的合并方法示例如图5.4所示。合并后的序列最简洁地反映了视频段中此区域的变化情况和质心 x 与 y 的变化情况。

```
合并前序列1：-1，0，-2，0，0，2，1，0，0，3，-2，0，-2，1，2，0
合并后序列1：-3，6，-4，3
合并前序列2：0，0，2，3，5，-1，0，-4，-5，0，0，-1，3，4，1
合并后序列2：10，-11，8
```

图5.4　序列合并方法示例

合并后的序列仍是时间序列，我们希望获取反映此序列特征的特征值。考虑到6项运动具有周期性，实验中计算出序列的两个主要特征值，一是序列的频率，序列长度与视频段时间的比值，反映宽窄变化频率；二是序列的时间平均方差，按式（5.1）计算。式中 X 是序列中任一值，式（5.1）表示序列中每个数值与整个序列的绝对值的平均值的差的平方和，再除以视频段时间，本书中称其为时间平均方差，请注意其与统计中方差的区别。以此时间平均方差反映序列中宽窄变化的时间平均幅度。

$$\text{时间平均方差}=\frac{\sum[X-E(|X|)]^2}{T} \tag{5.1}$$

如果横竖两个方向分别划分为5个区间，则一段视频最终可获取 $22\times2=44$ 个特征值。用这44个值组成一个特征向量，向量分量按横向1～5区间特征、纵向1～5区间特征、质心特征顺序排列，其结构如下：

$$[rf1,rd1,rsf1,rsd1,\cdots,cf1,cd1,csf1,csd1,\cdots,xf,xd,yf,yd]$$

其中，$rf1$ 表示第1行（横向第1区间）宽度变化的频率；$rd1$ 表示第1行宽度变化的时间平均方差；$rsf1$ 表示第1行空档变化的频繁；$rsd1$ 表示第1行空档变化的时间平均方差；$cf1$ 表示第1列高度变化的频率；……；xf 表示质心 X 变化的频率；……；yd 表示质心 Y 变化的时间平均方差。由此，提出的视频特征属于第5.3.1节所述的第2类特征。

4. 模式识别软件

PCP（Pattern Classification Program）是一组有监督学习模式识别分类算法开源软件，适用于由多维向量表征的模式分类任务[149]。PCP实现的算法有线性参数分类、二次参数分类、线性判别式分类、k 最近邻分类、神经网络分类以及支持向量机分类等，能够进行分类、交叉验证和主成分分析等。利用PCP进行分类识别，一般应先将数据集分成训练集和测试集两个子集；选择一种识别方法后一般先选择学习，以从训练集中学习到模式；在学习过程中有些算法的参数要确定和优化，PCP可以自动将训练集进行交叉检验（cross-validate）得到一些优化的参数，另外有一些参数需要用户指定；学习得到的模式用来对测试集进行识别，得出识别精度及详细的识别结果。识别精度的计算方法如式（5.2）。

$$\text{识别精度}=\frac{\text{识别正确的记录数}}{\text{测试集的总记录数}} \tag{5.2}$$

线性判别式模式识别方法设定一组线性判别函数，并利用训练样本计算线性判别函数的有关参数。PCP采用的是标准的最小二乘线性判别式分类算法。PCP的线性参数分类方法和二次参数分类方法采用的是正态分布的贝叶斯分类器。PCP的支持向量机使用的是台湾大学林智仁副教授等开发设计的SVM

模式识别与回归软件，Libsvm 的程序[150]。

5. 实验项目设计

为验证提出特征对行为识别的有效性，分别做了5组实验：

(1) 以一组特征数据为源，分别采用各类模式识别方法，测试数据的特性。

(2) 分别采用5×5、10×10、15×15、20×20划分前景，测试划分粗细程度对识别精度的影响。

(3) 考虑采用不同长度的视频分段对测试识别精度的影响。

(4) 对特征数据集进行线性判别分析，测试特征数据的线性可分性能。

(5) 利用第3.2节的特异数据挖掘算法测试数据集的类内与类间距离特性。

以下第5.3.3～5.3.6小节将分别阐述这5组实验内容、结果及结论，最后对整个实验进行总结。

5.3.3 数据分类特性测试

为了获得一定的精度，先选择了较精细的划分，在横竖两个方向均划分为20个区间，按第5.3.2节的方法对筛选出的视频段进行特征提取，共获得640行向量，每个向量的维数有 $(20\times2+20\times2+2)\times2=164$ 个。为从不同角度测试数据的有效性，分别对此640×164数据集进行了交叉检验、特征分级及降维后的交叉检验测试。

1. 不同识别方法的测试

为了测试模式识别方法在此数据集上的识别精度，一般的方法是将此数据集随机划分为训练集与测试集进行分类测试，实验选择了交叉检验。选择交叉检验，PCP程序将整个数据集随机分成 K 个子集，轮流取出一个子集做测试集，同时以另外 $K-1$ 个子集做训练集，测试分类精度。实验中选择了2子集交叉，做10次检验，取精度的平均值。PCP每次均将数据集随机分成2个子集，轮流作为训练集与测试集，得到2个测试精度，10次均值实际上是20个精度值的平均，这个均值的有效性更好。以不同方法进行交叉检验结果如表5.1所示。

表5.1　　以不同识别方法交叉检验结果

序号	识别方法	主要参数	平均识别精度
1	线性判别式分类器（Linear Discriminants）	无	92.1%
2	支持向量机（Support Vector Machine）	RBF核	35.8%
3	支持向量机	Linear核	95.6%
4	支持向量机	Polynomial核	95%

续表

序号	识别方法	主要参数	平均识别精度
5	支持向量机	Sigmoid 核	35.8%
6	k 最近邻分类器（kNN，k - Nearest Neighbor）	近邻数为 1，Euclidean 距离	93%
7	线性参数分类器（Linear Parametric Classifier）	无	93.8%
8	多层感知器（神经网络）（Multi - layer Perception）	1 个隐层，输入 164 节点，输出 6 节点，隐层 50 节点	77.5%（98.3%）

由表 5.1 可知，不同的参数或不同的识别方法对同一数据集交叉检验，结果差距很大；支持向量机分类时，其核函数的选择，目前国际上还没有形成统一的模式，一般凭经验或实验对比来选择。从这个实验中看出，采用不同的核，分类精度差距很大；多层感知器实际上应用的是神经网络方法，因为神经网络的学习过程是迭代过程，前几次的精度很低，表 5.1 中的 77.5%是全部迭代的平均值，所以是不准确的。如果以每次实验的最后两次迭代精度进行平均，其精度可达 98.3%。

表 5.1 显示的是总的分类精度，为了区分此特征对每类行为的表征能力，表 5.2 列出了 3 种分类器下各类别的召回率 R，R 的计算方法如式（4.9）。

表 5.2　　3 种分类器下各类别的召回率

召回率 R / 类别	各类记录数	线性判别式分类	支持向量机（Linear 核）	最近邻（$k=1$）
walking	36	0.833	0.889	0.944
running	46	0.478	0.848	0.500
jogging	44	0.727	0.909	0.773
handwaving	229	0.987	1.000	0.987
handclapping	131	0.969	0.939	0.885
boxing	154	0.948	0.981	0.942

表 5.2 显示，在用线性判别式分类器和最近邻分类器进行分类时，running 和 jogging 两类召回率很低，其他 4 类一直保持较高的召回率。其主要原因是跑和慢跑均为跑，实际应为一快一慢，但实验录像中两类的界限较模糊，这一点在文献［144］中也有说明。实验中，如果除去其中的一类记录，那么总体精度均提高了。

交叉检验实验说明，利用特征数据能较好地区分各类行为。

2. 特征值有效性分析

本部分进行测试的特征向量有 164 维，反映的是人体横向 20 个区间、纵

向20个区间、质心 x 和质心 y 的变化的频率和幅度，在分类中，这164个特征值所起的作用是否相同？哪些部分的分类特征较明显些？为了说明这个问题，利用640×164的数据集进行了特征选择分析。特征选择是指从一组特征中挑选出对分类最有利的特征，一般特征选择的目的是降低空间维数，本实验的目的是为了分析各分量对识别运动行为的支持度。选择PCP特征选择中的特征分级功能（feature ranking），并选择Euclidean距离作为判据进行按距离特征分级；PCP给出向量中各分量在分类中所起作用的等级，将等级值映射到［0，1］区间，得到164个反映分类等级的值。表5.3列出了特征分级等级排前40个的下标值及对应的等级。因为结果庞大，不方便以表格形式全部列出，以点分布形式显示为如图5.5所示。图5.5中横坐标表示向量中特征值的下标1～164，纵坐标表示对应特征的分类有效性等级0～1，其值越大，表示在分类中所起的作用越大。

表5.3　　特征分级等级排前40个的下标值及对应的等级

等级排名	下标值	等级值	等级排名	下标值	等级值
1	158	1	21	150	0.732004
2	116	0.903558	22	98	0.73181
3	82	0.87399	23	100	0.724216
4	120	0.865739	24	86	0.71268
5	112	0.848864	25	2	0.709577
6	108	0.840591	26	132	0.706989
7	124	0.836484	27	146	0.68965
8	114	0.835293	28	142	0.680631
9	104	0.827303	29	94	0.677147
10	154	0.82416	30	90	0.663689
11	128	0.822106	31	96	0.661202
12	118	0.822023	32	136	0.640245
13	110	0.81368	33	6	0.635413
14	122	0.802249	34	78	0.627403
15	130	0.800897	35	10	0.625385
16	126	0.794868	36	30	0.615844
17	134	0.782913	37	34	0.60155
18	106	0.771658	38	38	0.591807
19	102	0.764289	39	92	0.588769
20	138	0.732415	40	140	0.584262

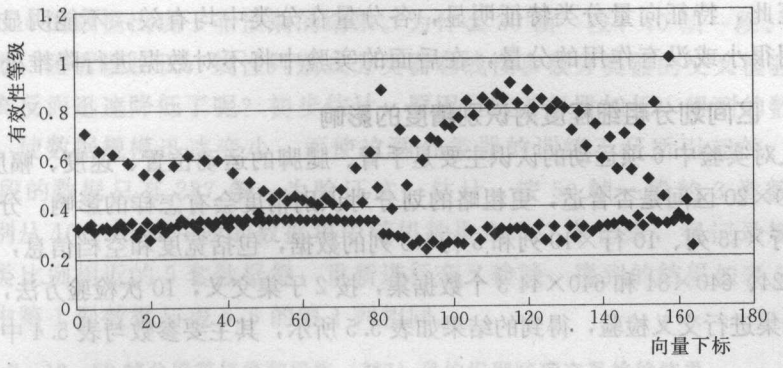


图 5.5 特征分级结果

由表 5.3 和图 5.5 可以得出，有效性等级大于 0.5 的，其下标值均为偶数，说明大部分时间平均方差值在分类中所起作用明显。另外，全部值的有效性等级均大于 0.2，说明所有特征均在分类中起到相当的作用，不能忽视。横向的 1～20 行特征反映在向量的 1～80 下标的特征值，其两端和中间作用明显，也就是头、手、脚部位的宽度变化在分类中作用显著；纵向的 1～20 列特征反映在向量的 81～160 下标的特征值，仍是两端和中间作用明显；161～164 特征反映的是质心 x 和质心 y 的波动，4 个值的分类作用均较小，这符合人对这 6 项运动的认识。

按第 5.3.2 节向量分量排列顺序，分量下标为 k 的特征，如果 k 或 $k+1$ 能被 4 整除，说明其反映的是空档变化情况；不符合这个特征，说明其反映的是宽度变化情况。将结果按分级值降序排序，发现对应下标有或没有“k 或 $k+1$能被 4 整除”特征的交杂在一起，说明空档的变化情况和宽度变化情况在分类中所起的作用是相当的。在横竖两个方向均划分为 20 个区间，按第 5.3.2 节方法对筛选出的视频段进行特征提取，但只提取宽度而忽略宽度内空档的数据，仍可获得 640 行向量，每个向量的维数有 $(20+20+2)\times 2=84$ 个。对此 640×84 数据集进行 2 子集交叉，10 交叉检验，测试结果如表 5.4 所示，与表 5.1 的 1、4、6、7 行对比，精度大大降低了。

表 5.4 不提取宽度内空档信息的数据集交叉检验结果

序号	识别方法	主要参数	平均识别精度
1	线性判别式分类器	无	89%
2	支持向量机	Polynomial 核	86%
3	k 最近邻分类器	近邻数为 1，Euclidean 距离	80.7%
4	线性参数分类器	无	90.9%

至此，特征向量分类特征明显，各分量在分类中均有效，不能明显找出分类作用很小或没有作用的分量，在后面的实验中将不对数据进行降维处理。

5.3.4 区间划分粗细程度对识别精度的影响

人对实验中6项运动的认识主要是手臂、腿脚的运动位置、速度、幅度不同。划分20×20区间是否合适，更粗略的划分对识别精度会有怎样的影响。分别提取了15行×15列、10行×10列和5行×5列的数据，包括宽度和空档信息，获得了640×124、640×84和640×44 3个数据集，按2子集交叉，10次检验方法，对此3个数据集进行交叉检验，得到的结果如表5.5所示，其主要参数与表5.4中相同。

表5.5 15行×15列、10行×10列和5行×5列数据集平均识别精度交叉检验结果

序号	识别方法	15行×15列	10行×10列	5行×5列
1	线性判别式分类器	93%	93.3%	92.3%
2	支持向量机	93.9%	92.9%	88.9%
3	k最近邻分类器	81.9%	92%	88.8%
4	线性参数分类器	93.9%	94.2%	93.6%

联合对比表5.1的1、4、6、7行与表5.5的结果，在区间划分取10行×10列以上时，可以获得较好的精度，更精细的划分获得的精度提高有限，但划分5行×5列的数据识别精度明显降低了。

5.3.5 视频分段对识别精度的影响

实验识别的6项运动具有周期性，对一段周期性序列，序列的分段会影响其频率和时间平均方差，一般分段影响会随着序列加长而减弱。“走、慢跑、跑”视频有自然的分段，而且每段均很短，无法再划分，所以选择“拳击、挥手、拍手”3类视频，进行不同的分段，以分析分段对此视频特征的影响。前面实验已将“拳击、挥手、拍手”3类视频按每30帧一段进行划分，获得514×164特征数据集，再按10帧一段、20帧一段、40帧一段和50帧一段进行段划分，横纵方向按20×20划分区间，采集宽度与空档变化数据，获得的数据集大小分别为1622×164、800×164、372×164、287×164。按2子集交叉，10次检验方法，对此5个数据集进行交叉检验，得到的结果如表5.6所示，其主要参数与表5.4中相同。

表5.6 10～50帧分段数据集平均识别精度交叉检验结果

序号	识别方法	50帧每段	40帧每段	30帧每段	20帧每段	10帧每段
1	线性判别式分类器	50.3%	80.2%	95.1%	96.7%	91.8%
2	支持向量机	98.6%	98.4%	97.6%	97.4%	92.3%
3	k最近邻分类器	97.5%	98.5%	97.7%	96.6%	90.8%
4	线性参数分类器	37.8%	79.3%	95.4%	96.9%	91.9%

表 5.6 的数据出现了非预期的结果。为什么 30 帧一段、40 帧一段、50 帧一段数据，随着段加长，线性判别式分类器和线性参数分类器的交叉检验精度没有提高反而迅速降低了呢？初步估计，原因是随着每段加长，得到的数据段数变少，使数据规模迅速变小，而使这两分类器的训练不足所引起的。由于 50 帧一段的数据只有 287 条，为验证这一估计，按 50 帧一段的 3 类数据比例，分别从 10～40 帧一段的数据集中随机抽取出 287 条记录，得记录规模相同、各类比例相近的 5 套数据集，重新进行交叉检验，得到的结果如表 5.7 所示。其中第 1 列数据与表 5.6 的第 1 列相同。

表 5.7 10～50 帧分段等规模数据集（287）平均识别精度交叉检验结果

序号	识别方法	50 帧每段	40 帧每段	30 帧每段	20 帧每段	10 帧每段
1	线性判别式分类器	50.3%	46.6%	44.4%	42.8%	39.1%
2	支持向量机	98.6%	97.6%	97.5%	94.8%	83.1%
3	k 最近邻分类器	97.5%	97.8%	95.2%	94.2%	80.9%
4	线性参数分类器	37.8%	35.0%	35.0%	34.4%	32.5%

由表 5.7，结果与预期相同，同时也看到了线性判别式分类器和线性参数分类器对训练集规模的敏感程度远高于支持向量机和 KNN 分类器。

再采用表 5.7 中的 4 种方法，其主要参数与表 5.4 中相同，利用 10～50 帧一段数据集为训练集，用 30 帧一段的数据集作为测试集，测试识别精度。为提高可比性，数据集规模均为 287，其结果如表 5.8 所示，表中“10～30 帧”表示以 10 帧每段的数据集作为训练集，以 30 帧每段的数据集为测试集的测试结果，其他类推。

表 5.8 利用不同段长训练集测试 30 帧一段数据集的识别精度的结果

序号	识别方法	50～30 帧	40～30 帧	30～30 帧	20～30 帧	10～30 帧
1	线性判别式分类器	98.9%	99.3%	100%	11.1%	15.0%
2	支持向量机	99.6%	99.6%	100%	60.6%	70.4%
3	k 最近邻分类器	100%	100%	100%	86.4%	77.7%
4	线性参数分类器	98.9%	98.6%	100%	10.8%	14.3%

其中，“30～30 帧”训练集和测试集相同，其平均识别精度均为 100%，这在一定程度上说明了数据集中各类数据的区分度较好；当分段长度达到 30 帧（约 1.2s）以上时，测试分类精度均超过了 98.6%，超过此长度，其分段长度的变化对分类精度的影响就很小了。

5.3.6 特征数据集其他分类性能分析与测试

前述测试主要是对利用特征数据进行分类精度对比，由于数据集大小、类

分布不同等因素影响，精度只能在一定程度上代表数据对分类的支持情况。为探索特征数据的性能，更深入的分析是必要的。作为分类的特征数据，希望其类间距离大而类内方差小，也就是说，不同类别间的特征值距离较远，而同一类别内的特征距离较近。为分析特征数据的距离特性，分别进行了判别式分析和特异分析。

1. 判别式分析

为衡量特征数据间的距离，常规的就是利用欧氏距离，欧氏距离也适合本节数据集的计算。鉴于第5.3.5节的分类方法中，线性判别式分类器分类精度较好，选择利用此分类器中的线性判别式来计算数据集的距离特征。前面的实验表明，在区间划分取10行×10列以上时，可以获得较好的精度，为了不增加计算的复杂度，选择10×10划分的数据集，按6类数据规模大致相同的比例，组成一组新数据集，其记录构成为：walking－36、running－40、jogging－40、handwaving－40、handclapping－40、boxing－40，数据集大小为236×84。利用SPSS将集合进行判别式分析，分析以全部84维向量为独立的变量，建立Fisher线性的判别式，并根据判别函数来计算各类间距离与类内距离。

基本的Fisher判别方法是一种两类别判别方法，它利用使Fisher准则达到最大值的方向作为最优投影方向，样本模式在该方向投影后的类间散度达到最大而类内散度达到最小。以投影函数作为判决函数function，当样本的function的得值大于某个阈值时判定为一类，否则判定为另一类。判决阈值f_0的典型选择有3种[94]：

$$f_0=\frac{1}{2}(\tilde{u}_1+\tilde{u}_2) \tag{5.3}$$

$$f_0=\frac{1}{N_1+N_2}(N_1\tilde{u}_1+N_2\tilde{u}_2) \tag{5.4}$$

$$f_0=\frac{1}{2}(\tilde{u}_1+\tilde{u}_2)+\frac{1}{N_1+N_2-2}\ln\frac{P(w_1)}{P(w_2)} \tag{5.5}$$

式中：$\tilde{u}_1$、$\tilde{u}_2$为function作用在第1类和第2类样本上的平均值；N_1、N_2为第1类和第2类样本的数量；$P(w_1)$、$P(w_2)$为第1类和第2类样本先验概率。

对于多类的判别问题，以一类为判别目标，其他样本均设定为另一类，可以构造一判别函数；在剩余的类中，再以一类为判别目标，其他样本均设定为另一类，构造第2个判别函数；重复此过程，一直到每一类均能判别。所以一般n类分类问题需构造$n-1$个判别函数。

利用一Fisher判别函数，可以计算出每个记录的得值y。如果将记录数为n的数据集按其类别分为k组，第i组的记录数为n_i，那么所在组间距离的平

方和也称为组间散度，用式（5.6）表示：

$$SSA=\sum_{i=1}^{k} n_i(\bar{y}_i-\bar{y})^2 \tag{5.6}$$

式中：$\bar{y}$ 为全部数据的判别式得值的平均值；$\bar{y}_i$ 为第 i 组数据得值的平均值。

可见，SSA 是各组平均值与总体平均值离差的平方和，反映了组间的总距离。SSE 则反映了组内离差平方和，也称组内散度，其计算方法如式（5.7）所示：

$$SSE=\sum_{i=1}^{k}\sum_{j=1}^{ni}(y_{ij}-\bar{y}_i)^2 \tag{5.7}$$

F 则是平均组间平方和与平均组内平方和之比，其值可有效表征数据集中组间松散与组内紧密的对比，计算方法如式（5.8）所示：

$$F=\frac{SSA/(k-1)}{SSE/(n-k)} \tag{5.8}$$

式（5.8）的 F 服从（$k-1$，$n-k$）个自由度的 F 分布，根据 F 分布表可以计算出其相伴概率值。如果相伴概率值小于显著性水平 a，则认为各组间总体均值有显著差异[151]。

因为数据集中有 6 类数据，建立了 5 个判别式 Function 1～Function 5，由各判别式的 Structure Matrix 可知，特征向量的不同分量与不同的 Function 显著相关，各判别式的特征值如表 5.9 所示。

表 5.9　　5 个判别式的特征值

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	41.123	65.5	65.5	0.988
2	10.703	17.1	82.6	0.956
3	4.73	7.5	90.1	0.909
4	4.195	6.7	96.8	0.899
5	2.007	3.2	100.0	0.817

其中的特征值 Eigenvalue 即为式（5.5）中的 F 值，表 5.9 的第 1 行表示利用 Function 1，计算出的 F 值为 41.123，它对整体分类的贡献是 65.5%，下一列是累计的贡献百分比，最后一列为典型相关系数，反映的是此判别函数与组别间的关联程度。可以看出 5 个判别式累计分类的贡献率是 100%，与分类类别相关程度均大于 0.8；由 Structure Matrix 可知，数据集中的 84 维特征分别与 5 个判别式函数达到相关性显著水平；表 5.10 中 1～6 类代表前述的 6 类行为：walking、running、jogging、handwaving、handclapping、boxing，

各类数据在不同的判别式的中心有显著差距，说明 5 个判别式对此数据集的分类判别是有效的。

表 5.10　　各类记录在 5 个判别式上的中心

类别	Function				
	1	2	3	4	5
1	−5.779	0.952	3.668	−1.151	−1.650
2	−7.091	2.415	−3.258	1.303	−0.849
3	−5.225	−0.382	0.704	−0.360	2.820
4	10.028	4.922	0.252	−0.179	0.301
5	3.412	−3.850	−1.807	−3.107	−0.481
6	4.076	−3.962	0.808	3.378	−0.305

如图 5.6 所示显示出由 Function 1 和 Function 2 计算出的各类数据的分布情况，其中，handwaving 类别（4），已经能有效区分，但其他类别还要进一步识别。

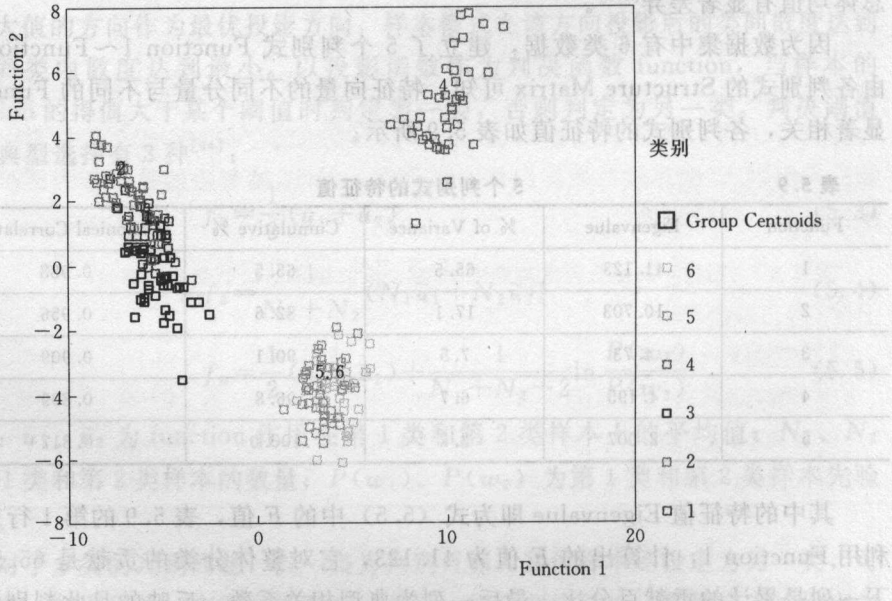


图 5.6　由 Function 1 和 Function 2 计算的各类数据分布图

如表 5.11 所示列出了最终的分类结果，可以看出分类错误仍出现在 running 和 jogging 之间，这和前面的测试是一致的。

这组分析表明，特征数据集具有较好的线性可分特性。

表 5.11 Fisher 线性判别的分类结果

原类别		Predicted Group Membership						Total
		1	2	3	4	5	6	
Count	1	36	0	0	0	0	0	36
	2	0	39	1	0	0	0	40
	3	0	0	40	0	0	0	40
	4	0	0	0	40	0	0	40
	5	0	0	0	0	40	0	40
	6	0	0	0	0	0	40	40
%	1	100.0	0.0	0.0	0.0	0.0	0.0	100.0
	2	0.0	97.5	2.5	0.0	0.0	0.0	100.0
	3	0.0	0.0	100.0	0.0	0.0	0.0	100.0
	4	0.0	0.0	0.0	100.0	0.0	0.0	100.0
	5	0.0	0.0	0.0	0.0	100.0	0.0	100.0
	6	0.0	0.0	0.0	0.0	0.0	100.0	100.0

2. 特异分析

在数据集中，一些数据或对象与其中其他数据或对象显著不同，则称是特异数据或特异对象。同一类特征数据相似性高，而不同类数据相似性低，在一类数据中掺入少量的其他类数据，希望掺入的数据能被识别为特异的。选择 10×10 划分的数据集，构造 4 个子集如表 5.12 所示，其中每个子集均有一大类，另外的类别记录加一起为小类，约占总记录的 10%。

表 5.12 4 个子集组成表

子集名称	walking	running	jogging	handwaving	handclapping	boxing	总记录数
Subt1	5	5	5	229	5	5	254
Subt2	3	3	3	3	3	154	169
Subt3		46	2	2		2	52
Subt4	2	46			2	2	52

以 4 个子集为数据，以大类为正类，以小类为负类，利用本书第 3.2 节的全局特异数据挖掘算法，计算记录的特异因子，将其排序后绘制 ROC 图，结果如图 5.7 和图 5.8 所示，各 ROC 曲线下面积值如表 5.13 所示。

表 5.13 各 ROC 曲线下面积

曲线	Subt1	Subt2	Subt3	Subt4
面积	0.864	0.930	0.678	0.732

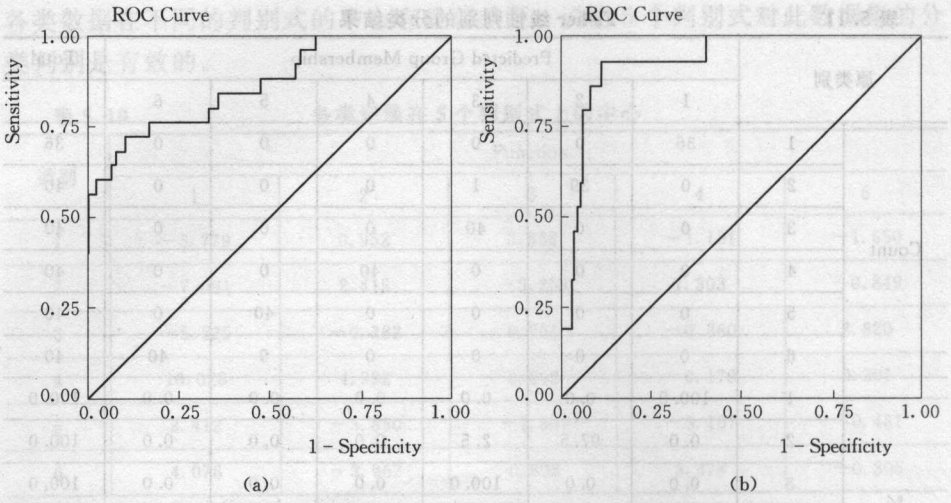


图 5.7　Subt1 和 Subt2 的特异因子 ROC 图

(a) Subt1；(b) Subt2

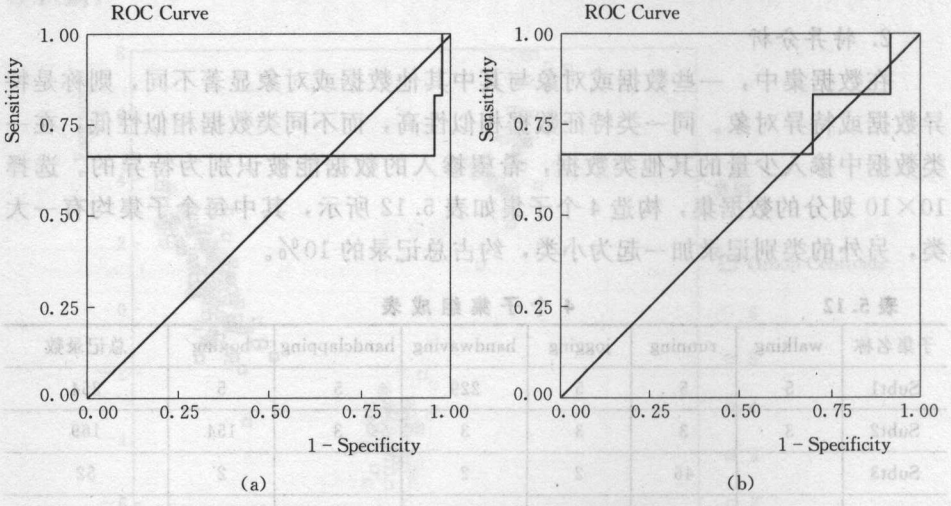


图 5.8　Subt3 和 Subt4 的特异因子 ROC 图

(a) Subt3；(b) Subt4

由以上的图表可反映出，小类数据均表现出了较好的特异性。在 Subt3 的 ROC 曲线性能是最差的，主要原因是其中涉及 running 和 jogging 的区分。

至此，本节提出的行为识别特征数据具有较好的“不同类别间的特征值距离较远，而同一类别内的特征距离较近”的特性。

5.3.7 实验验证总结

从感观上，人的运动在2D视频中表现在身体不同部位的伸与缩及期间空档的伸与缩。将人运动前景矩形在横和竖方向上划分为若干区间，采集运动过程中各区间宽度与空档伸缩变化情况的序列，采集运动过程中质心X和Y变化的序列。计算这些序列的频率和时间平均方差形成特征向量，利用此特征向量数据进行了多方面的实验，验证其在人的行为识别方面的有效性。实验说明：

(1) 特征向量易于提取，向量分类特征明显，获得了高的实验精度；各分量在分类中均有效，不能明显找出分类作用很小或没有作用的分量。

(2) 区间划分的粗细程度对识别精度有明显的影响，一般区间划分取10行×10列以上，可以获得较好的精度。

(3) 因为特征数据表征的是视频序列的伸缩频率与时间平均方差，视频段的大小、起始帧的位置均会影响特征数据。实验表明，当分段长度达到30帧以上时，就可获得很高的分类精度，且分段长度的变化对识别精度影响就会很小了。

(4) 行为识别特征数据具有较好的“不同类别间的特征值距离较远，而同一类别内的特征距离较近”特性。

实验能说明提出的视频特征对行为识别的有效性，但实验还较粗糙，要使特征数据在现实场景中应用，更细致深入的实验、分析与变换是必要的：

(1) 由于前景提取技术粗糙，实验中只能采用Schuldt视频库中的部分视频，实验识别精度与采用同样视频库的文献中的精度[144]不可比。

(2) 区间划分与段划分粗略，所以“区间划分取10行×10列以上”和“分段长度达到30帧以上”不是精确的边界值。

(3) 特征的提取基于运动中的整个单人人形，在多人场景和人体被部分遮蔽时特征会失效；前3种运动主要是侧面的录像，后3种运动主要是正面的录像，特征数据不具有视角不变性。特征数据深入的变换还需研究。

5.4 小结

分析了智能监控技术现状，提出一智能监控系统的构架。系统数据由3层构成，即原始录像的视频文件、模式和实时数据。模式主要有特征模式、行为模式、异常模式；实时数据有特征数据与行为数据。从原始录像中提取数据形成特征模式，由特征识别行为。

提出一种行为识别的视频特征。观察人运动的2D视频，不同的运动行为在一定程度上表现为人身体不同部位的伸与缩。将人运动前景矩形在横和纵方

向上划分为均匀的区间，采集这些区间的宽度及其内部空档变化的序列，以序列的频率和时间平均方差构成特征向量。为了验证此特征对行为识别的有效性，采用线性判别式方法、支持向量机方法、k 最近邻方法、线性参数分类方法等模式识别方法，进行了分类交叉检验、特征值分析，进行了不同粗细划分的特征数据识别精度对比，进行了不同视频分段的识别精度对比，进行了线性划分和特异因子计算实验。实验结果表明，当视频分段长度达到一定值，区间划分达到一定精细程度时，利用特征数据能有效识别不同的行为，特征数据线性可分性较好，具有较好的“不同类别间的特征值距离较远，而同一类别内的特征距离较近”的特性，并且特征向量的各分量在分类中均有效。提出的特征直观、易于获取、对镜头远近有鲁棒性，避免了识别与跟踪人身体各部分的困难。

第6章　基于差分的行为特征与基于全前景的行为特征比较

视频中的底层特征是提取行为识别的重要环节。提出一个基于表观的特征集，分别从视频帧差序列和视频前景的全景序列提取特征，利用隐马尔可夫模型（HMM）来对特征的时间变化建模，并分别利用特征选择和投票方法，对两类特征的性能进行了对比。实验结果表明提出的特征集对行为识别有效。同时，基于差分序列提取的特征对行为识别的能力略好于从前景全景序列提取的特征的识别效果。让大家看到了以差分序列为基础，识别行为的可行性。差分序列相比较全景序列具有对噪声不敏感的特征，且差分提取的特征简单、计算复杂性小。这验证了本章工作的意义。

6.1　概述

在过去十几年中，很多研究者把兴趣集中在2D视频中的行为分析上，而且受现实应用的复杂场景驱动，这些研究工作仍然在深入进行[154-157]。视频中的行为分析技术应用范围包括智能监控、基于内容的视频存储和检索、交互的虚拟现实系统、智能交通、智能建筑等。尽管很多人的行为分析的算法已经被开发出来，它们的效果在复杂应用中还有待提高。

通常，视频中人的行为分析有两类主题：分析视频中单个人或几个人的行为；分析视频中群体行为、人群流动路径等。本章关注的是前者。视频中个人的行为分析框架一般包含特征提取、行为建模和行为识别几个过程。目前对视频分析方法的分类主要基于方法所采取的视频特征。典型的方法分为：基于表观特征的方法、基于兴趣点的方法、基于身体部位特征的方法，以及基于运动的方法等。

数字视频是数字图像的序列，多数研究从整个图像序列或者整个前景序列中提取特征。Bobick和Davis首次尝试从视频的差分序列中提取特征识别行为[153]。行为识别依据从视频帧差序列建立的运动能量图（Motion Energy Iamge，MEI）和运动历史图（Motion History Image，MHI）。之后，出现了若干基于MEI和MHI的行为分析工作。Rougier等利用MHI可以表示运动历史的特性，提出了一种联合MHI和人形检测跌倒的方法[158]。Murayama等

提出了一种基于 MHI 梯度的方向识别行为的方法。他们相信 MHI 梯度的方向很好地表达了运动行为[159]。Han 等提出了步态能量图（Gait Energy Image，GEI）来描述人的走路步态[160]。之后，为了更好地表达步态的时间信息，Chen 等将 GEI 扩展到多个通道图像上[161]。Javed 等从全景前景序列中提取 MHI 特征来识别运动[162]。事实上，除了 Rougier 和 Murayame 的工作，其他研究中的能量图都是从全景序列中提取的。这是否是因为其他研究者不能从差分序列中提取出足够的信息？从差分序列中提取的特征与从全景序列中提取的特征对行为的表征能力的差距到底有多大呢？本章提出一个基于表观的特征集，分别从差分序列中和从全前景序列中提取特征，并利用特征选择和特征投票的方法进行行为识别。以从差分序列提取特征、投票方法识别为例。首先，从视频段中计算差分序列；从差分序列中提取特征；选择特征，形成特征子集，利用 HMM 进行行为建模，并估计各特征子集对识别行为的贡献。之后，对于一段新的视频，同样的特征被提取，利用学习得到的 HMM 进行识别。最终的行为，将由各模型识别结果投票来判定。整体过程描述如图 6.1 所示。

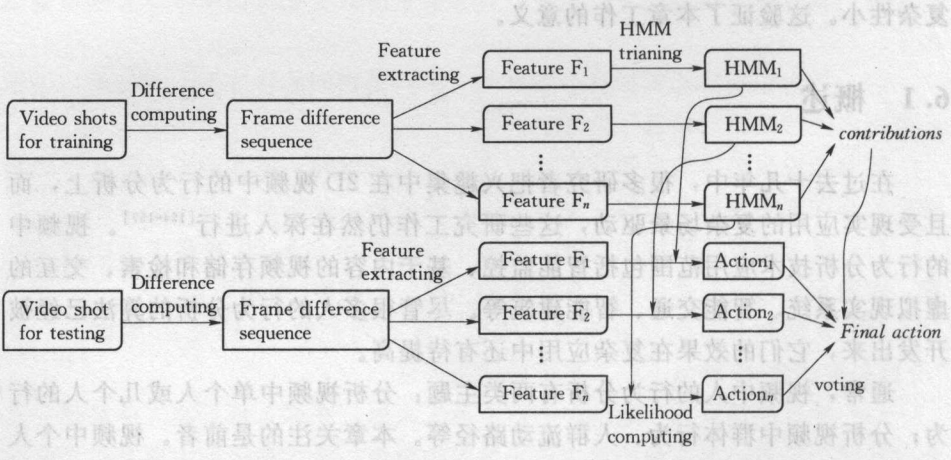


图 6.1　行为建模与识别框架图

第 6.2 节总结现有的基于表观特征进行行为识别的研究工作；第 6.3 节对比差分序列与全前景序列；第 6.4 节详细描述本章利用的特征集中的特征；最后进行实验分析和总结。

6.2　表观特征

表观特征在行为识别中应用很早。区域和轮廓是最直接的表观特征。对于一帧图像，从前景区域和轮廓中提取的特征可以表示姿态或姿势。一些时间过

程的模型可以用来对行为建模，如：HMM，动态时间弯折等[163,164]。

减背景方法是获取前景区域最常用的方法。这种方法利用当前帧与背景图像的差分图像中具有较大差值的区域来获取前景。当然，应用此方法时背景图像的精确性非常关键。复杂场景中，基于高斯混合模型（Gaussian Mixture）的背景估计方法较有效[165,166]。获得了前景图像，之后的首要问题就是什么样的特征可以有效表达行为。Hota 等测试了监控视频中的特征，其测试表明，一些表观特征可以有效区分视频中的人和其他物体，如：Hu 不变矩（Hu invariant moments）、最小外接矩形（Minimum Bounding Rectangle，MBR）的高宽比、填充率（fill ratio）、周长、一块区域的紧凑性和凹凸性（compactness and convexity）、凸偏差（convex deviation），以及投影直方图（projection histogram）等[152]。Lee 等利用曲线演化技术来识别人的外形[167]。Jiang 等提出了一种匹配外形序列识别行为的方法[168]。本章的目标是对比差分序列中的表观特征和全前景序列中的表观特征对行为的表征能力。

6.3 帧差序列与全前景序列

经过时间差分，获取的图像转换成二值图像。在这个二值图像中只有变化的运动而不是全部运动可以被检测出来，这可能造成一些缺失。一些像颜色、纹理、身体形状的信息丢失了。差分图像很容易产生空洞，且当物体停止运动时，信息就消失了。差分序列可以在一些场合提供支持信息用于进一步的精确分割。一方面 Hu 等人利用差分来检测运动的区域，再利用高斯混合模型检测精确的目标[169]。Yan 等人利用差分图像中提取的参数，之后通过匹配肤色、体形以及姿势等来精确提取运动中人的形状[170]。另一方面，前述工作说明累积的差分形状 MEI 和 MHI，可以表达行为，本章关注的是差分序列中一帧一帧中提取的信息。

将包含运动的视频中连续两帧相减，会产生两类突出的点集，一类中包含大的正值，一类中包含大的负值。如果按绝对值阈值化，在差分图像中形成近似运动方向两边边缘的前景。如果只按一个点集阈值化，比如只将正值的点集阈值化而忽略负值点集，那么产生的二值图像可以近似看作运动方向单边边缘，相当于产生了半差分图像。如图 6.3 所示中显示的是“走”视频的相关图像。其中，第 1、第 2 是连续两帧原图像，第 3 帧是前两帧直接相减后获取的灰度图像，第 4 帧是阈值化后的差分图像，第 5 帧是半差分图像。视频来源于一个流行的行为分析数据库 Weizmann 人的行为数据库[171]。在它的主数据库中包含 10 种行为：弯腰（bending）、杰克跳（jumping jack）、跳（jumping）、立定跳（jumping in place，pjump）、跑（running）、侧跳（jumping sideways）、跳跑

(skipping)、走 (walking)、单臂挥 (one hand waving, wave1) 和双臂挥 (two hands waving, wave2)。帧速率是 25 帧/s，每帧是 144×188 大小。为比较方便，本章中的主要例子视频均来源于 Weizmann 数据库。这个数据库中提供每个视频的精确背景图像，所以采用减背景方法获取全前景图像。

如图 6.2 所示中显示了这个视频库中“弯腰”和“走”两段视频中提取的全前景图像序列、差分图像序列和半差分图像序列。其中，两个运动是“走”和“弯腰”。对于每个运动第 1 行显示的是全前景图像序列，第 2 行显示的是差分图像序列，第 3 行显示的是半差分图像序列。可以看到，差分图像近似地反映了运动中一个时刻的运动部分的两边轮廓。“弯腰”时，因为运动是原地的，在差分图像中只有头部和身体上半部分有运动的部分可以检测到。“走”时，因为落地的腿是静止的，所以在差分图像中消失了。半差分图像序列中因为只是近似了运动方向的单边缘，其中的信息更少了。人在差分图像或者半差分图像中识别行为可能有困难。为了减少前景重叠带来的困难，本章和下一章中将利用的是半差分图像序列，为了表达方便，并且不至于引起混淆，以下所提及的差分图像均指半差分图像。我们将证明，如果合适的特征提取出来，计算机算法也许可以区分它们。相邻帧意味着小的时间间隔，一般来说，光线和外形等变化最小，在复杂环境中鲁棒性好。另外，差分图像比全前景图像获取更容易，计算相对简单。所以，探索差分图像中提取行为特征的方法有实际意义。

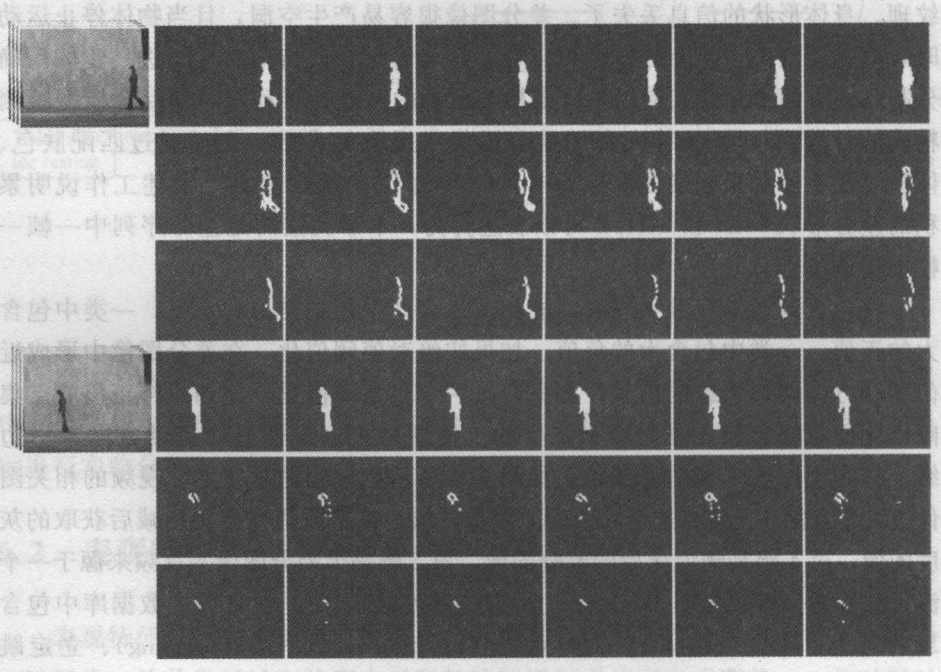


图 6.2　全前景图像序列、差分图像序列和半差分图像序列的示例

图 6.3 “走”视频的示例图像

6.4 特征集

从差分序列图像或者从全前景图像中提取若干基于表观的特征，这些提取的特征接下来会用来比较对行为识别的有效性，也会被用于投票方法的比较中。这些特征最终被表示为 31 值的向量 $M=(m_1, m_2, \ldots, m_{31})$，本节接下来详细描述。这些值被规划为 10 个类，表示为 $(F_1, F_2, \ldots, F_{10})$。为了减少镜头远近的影响，这些特征将利用正立的最小外接矩形（Upright Minimum Bounding Rectangle，UMBR）的参数进行标准化。UMBR 是一个内接全部前景区域并且与坐标轴平行的矩形，请注意与前述的 MBR 不同[152]。本章中利用 UMBR 进行标准化是假设视频中人的行为是垂直于水平线进行的。

1. F_1：8 扇区的轮廓特征

轮廓是表观特征中表示形状的重要特征。一种表示闭区域外形凹凸描述符如图 6.4（a）所示[172]。描述符的值由形心到轮廓上的每个点的距离确定，图中带有箭头的线段表示从质心到边界上某一点的距离。为了避免维缩减过程，将 360°方向划分为 8 个扇区，如图 6.4（b）所示。图中坐标为行列坐标（col, row），原点设在前景区域的质心。每个扇区中，计算质心到边界的最大距离，图中显示的 3 个带有箭头的线段表示第 2、第 3、第 4 扇区的最长距离。之后，这个距离要除以 UMBR 的长对角线长度，以进行标准化。这样获得了 $m_1 \sim m_8$，粗略表达运动时部位的伸缩。由于光线、遮蔽等影响，获取的全景图像，或者获取的差分图像中的前景区域有可能含多块分离的区域。在本章中，分离的块将作为整体进行质心、面积、UMBR 的计算。

2. F_2：8 扇区像素分布描述符

如图 6.4（b）所示中 8 扇区的像素计数用于描述 8 个方向的像素分布。设 A 是前景区域的面积，8 个方向计数值除以 A 来标准化，结果是 $m_9 \sim m_{16}$。

3. F_3：主轴方向

主轴可近似表达运动时身体的角度、重心和对称信息，且对噪声不敏感[173]。对于一幅前景图像，主轴可以由式（6.1）计算，其结果值表示前景主轴与水平向右的 x -轴逆时针方向的夹角。M_{11}、M_{02}和 M_{20}是中心矩。通过 $\pm\pi/2$ 使结果调整到 $0 \sim \pi$ 之间，这个值就是 m_{17}。如图 6.5 所示显示了 7 个图

像和它们前景的主轴，其主轴值分别是：1.6341、1.6276、1.6268、1.6124、1.5894、1.5853、1.5747。

$$\theta=\frac{1}{2}\arctan\left(\frac{2M_{11}}{M_{20}-M_{02}}\right)\pm\frac{\pi}{2} \tag{6.1}$$

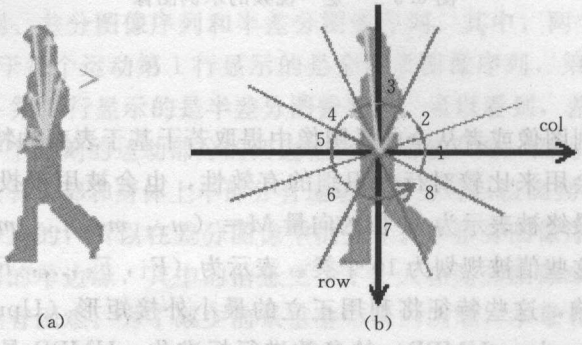


图 6.4　一种形状描述符

(a) 一种闭区域形状描述符；

(b) 将行列坐标原点设置在形心，方向划分成八扇区后的形状描述符

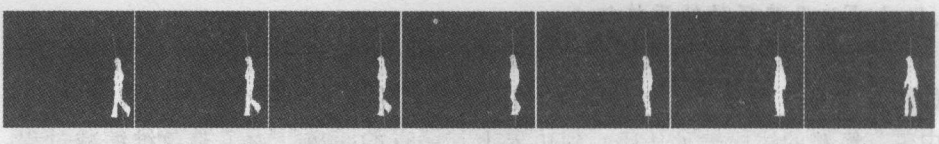


图 6.5　“走”的 7 个前景图像和它们的主轴

4. F_4：相对质心运动

视频中运动物体的路径提供位置和速度信息。在室外监控视频中的运动路径用以表达简单的事件，如一个人从入口 A 移动到入口 B[174]。利用视频帧之间前景质心的相对位置序列表示路径。因为第 1 帧之前没有位置信息，所以路径的初始值设为 (0，0)。相对位置是 (m_{18}，m_{19})。如图 6.6 所示显示了几个行为的运动路径。

5. F_5：Hu 矩

Hu 提出了 7 个对位置、缩放和旋转不变的矩的计算方法[175]。Hu 矩在目标识别中有较好效果。引用 Hu 的 7 个矩作为形状与前述特征不同的表示，它们是 $m_{20}\sim m_{26}$。

6. F_6，F_7，F_8，F_9，F_{10}：其他表观特征

依据 Hota 等人测试过的几个表观特征，设计出 F_6，F_7，F_8，F_9和 F_{10}，说明如下：

F_6，填充率。填充率是前景面积 A 与 UMBR 面积的比值，设为 m_{27}。

F_7，高宽比。是 UMBR 的高与宽的比值，设为 m_{28}。

F_8，紧密度（C）。表达前景区域的紧密程度，用 $C=\frac{P^2}{4\pi A}$ 计算。P 是前景区域的周长，设为 m_{29}。

F_9，域凸性（SC）。前景区域的周长与面积的平方根的比值，计算公式是 $SC=\frac{P}{\sqrt{A}}$。设为 m_{30}。

F_{10}，凸偏差（CD）。由 $CD=\arctan\left(\frac{1}{C\times SC}\right)$ 计算，设为 m_{31}。

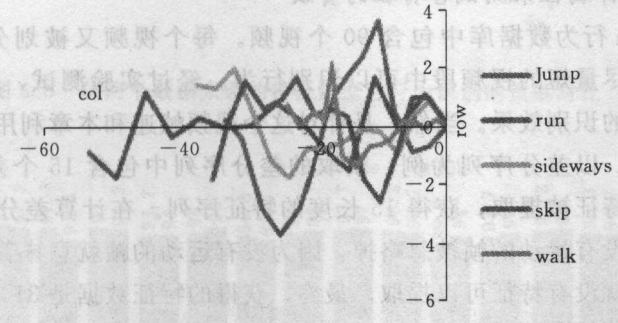


图 6.6 几个从右到左运动的行为的路径

6.5 实验分析

1. 建模与识别

马尔可夫链是一种描述时间序列变化规律的随机过程。它的限定条件是历史有限性，也就是当前状态只与前一个状态有关。隐马尔可夫模型 HMM 是一个符合包含隐含状态的马尔可夫链的统计模型。尽管人的运动不完全符合历史有限性，但很多研究都表明人的行为可以由 HMM 有效建模[163,164,176]。

一个 HMM 是 5 元组：$\mu=(S, O, \Pi, A, B)$。其中，S 和 O 是状态集合和观察集合，Π 是初始状态矩阵，A 是状态之间的转换概率矩阵，B 是从状态到观察的转换矩阵。利用具有同一行为标签的一组观察，通过最大化概率 $P(O\mid\mu)$ 可以获得模型 μ。利用各行为的观察数据学习得到的模型表示为 $\{\mu_1, \mu_2, \cdots, \mu_i, \cdots\}$。一个新的、没有行为标签的观察，计算它与各个 HMM 模型的似然度，将它分类为最大似然的类别，就是 $\text{argmax}P(O\mid\mu_i)$。利用 HMM 进行训练和识别行为的流程如图 6.7 所示。

本章和第 7 章中 HMM 的训练和识别利用的是 Kevin Murphy 的贝叶斯工

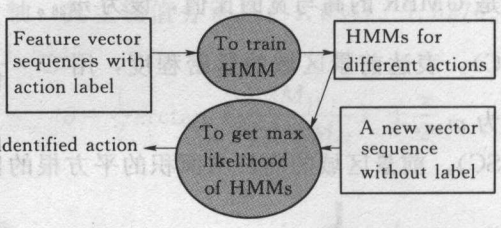


图 6.7　HMM 模型训练与利用 HMM 进行识别的流程图

具，其中包含了 HMM 的训练和似然度计算模块[176,177]。

2. 利用选择属性来测试各特征的贡献

Weizmann 行为数据库中包含 90 个视频。每个视频又被划分为视频段。大家都希望在尽量短的视频段中可以识别行为。经过实验测试，15 帧一段视频足够获取好的识别效果。当然，是针对这个视频帧速和本章利用的特征与建模方法来讲的。以差分序列为例，获取的差分序列中包含 15 个差分的图像。之后，前述的特征被提取，获得 15 长度的特征序列。在计算差分序列时，连续 3 帧或更多没有运动的帧被忽略掉，因为没有运动的帧就意味着差分图像中没有前景，也就没有特征可以提取。最终，获得的特征数据是 31×15 的序列，称差分序列的原始全特征序列。从全前景序列中提取的特征数据仍是 31×15 的序列，称全前景序列的原始全特征序列。作为比较，以下的测试将分别在这两类特征序列上进行。

测试利用留一策略（leave - one - out）：对于一种行为的特征序列集，随机抽取一个序列作为测试序列；利用其他的序列进行 HMM 训练，获得 HMM 模型；利用获得的模型对保留的测试序列进行测试；这个过程进行多次，计算测试精度。在 HMM 中，利用高斯混合模型来规划观察，因为特征是连续值。测试中，隐藏状态设为 2 个，高斯混合模型设为 3 个。这两个参数有变动时，测试精度有少量改变。每次测试为 10 个行为中的每个行为保留一特征序列作为测试序列，另外的序列用来训练 HMM。之后，对保留的 10 个序列进行测试，计算识别精度。

为了测试前述特征集中不同特征对行为识别的贡献程度，也为了对比从全前景序列中提取的特征数据与从差分序列中提取的特征数据对不同行为识别的效果，从两套原始全特征序列中进行了组合型特征选择，形成特征子集。利用两套特征子集进行行为识别测试，测试结果对比显示如图 6.8 和图 6.9 所示。其中，x 轴表示选择的特征子集的构成，y 轴表示利用相应子集测试后的精度对比。图中的每个精度值均为 30 次测试的平均精度，精度值的标注是百分比重。从两图中可以看出最高的识别精度达到

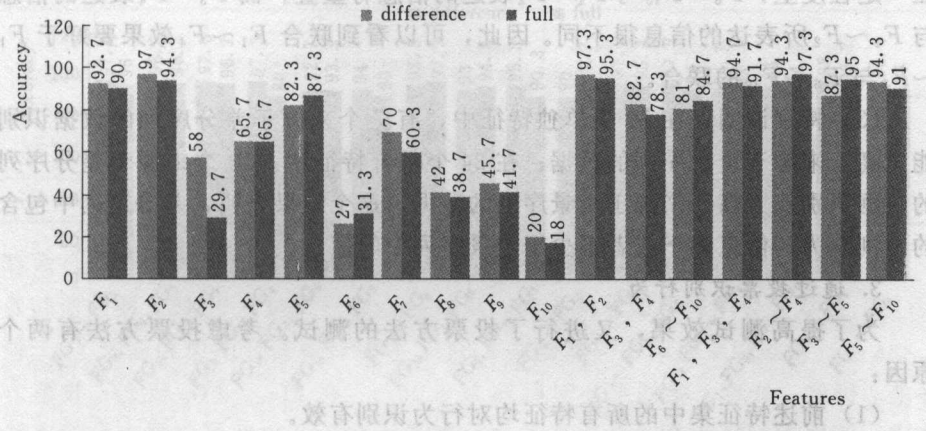


图 6.8 从两个数据来源的特征数据对行为识别的贡献对比（1）

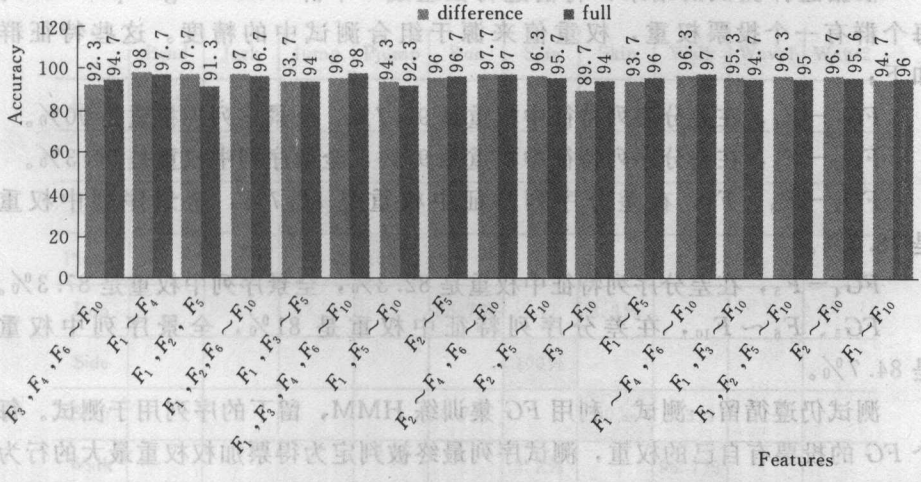


图 6.9 从两个数据来源的特征数据对行为识别的贡献对比（2）

了 98%。并且有：

（1）对于单项特征，F_1和 F_2对行为识别效果最好，表明具有方向的外观特征能更好地表征姿态。Hu 矩（F_5）也可以表达形状，但是它们是对位置、伸缩及旋转不变的，也就是说这些特征丢失了，所以只用 Hu 矩来识别行为效果不好。F_3、F_4、$F_6 \sim F_{10}$也可以表达形状，但是每个特征只表达了形状的一些方面，单独使用效果不好。

（2）通过联合使用多种特征，精度普遍提高了。应该说每种特征都对行为识别有贡献，但不是特征用得越多识别效果越好，联合全部特征并不能获得最好的精度。例如，差分序列中 F_2联合 F_3和 F_4后比只利用 F_2的精度更低了。

在一定程度上，$F_6 \sim F_{10}$与$F_1 \sim F_2$表达的信息有重叠，而$F_3 \sim F_4$表达的信息与$F_1 \sim F_2$所表达的信息很不同。因此，可以看到联合$F_1 \sim F_4$效果要好于$F_1 \sim F_2$与$F_6 \sim F_{10}$的联合。

(3) 特别说明，在10个单独特征中，有7个来源于差分序列的数据识别能力高于来源于全景序列的数据；在24个联合特征中，13个来源于差分序列的数据识别能力高于来源于全景序列的数据。这个结果表明，差分序列中包含的识别行为的信息多于或者不少于全景序列中的信息。

3. 通过投票识别行为

为了提高测试效果，又进行了投票方法的测试。考虑投票方法有两个原因：

(1) 前述特征集中的所有特征均对行为识别有效。

(2) 单纯联合更多的特征并不能提高识别精度。

根据选择测试的结果，将前述特征组成5个群 (feature groups，FGs)。每个群有一个投票权重，权重值来源于组合测试中的精度。这些特征群如下：

FG_1-F_1，在差分序列特征中权重是92.7%，全景序列中权重是90%。

FG_2-F_2，在差分序列特征中权重是97%，全景序列中权重是94.3%。

FG_3-F_3、F_4，在差分序列特征中权重是82.7%，全景序列中权重是78.3%。

FG_4-F_5，在差分序列特征中权重是82.3%，全景序列中权重是87.3%。

FG_5、$F_6 \sim F_{10}$，在差分序列特征中权重是81%，全景序列中权重是84.7%。

测试仍遵循留一测试。利用FG集训练HMM，留下的序列用于测试。每个FG的投票有自己的权重，测试序列最终被判定为得票加权权重最大的行为类别。如图6.10所示显示了FG投票测试的结果。图中每个精度是30次测试的平均值。

将图6.10与图6.8和图6.9对比，对于来源于差分序列的数据，16个精度中，9个投票方法获得的精度高于对应的组合方法获得的精度；对于来源于全景序列的数据，16个精度中，12个投票方法获得的精度高于对应的组合方法获得的精度。如图6.10所示的16个精度，12个基于差分序列的数据测试精度高于基于全前景序列的数据测试精度。

FG_1、FG_2和FG_4基于差分序列特征数据识别精度均达到98%。如图6.11所示中显示了这个测试的混淆矩阵，以进一步观察识别效果。可以看到，其中running、jumping sideways、skipping和walking有些混淆。观察这些行为的差分序列，可以看到它们确是相似的。

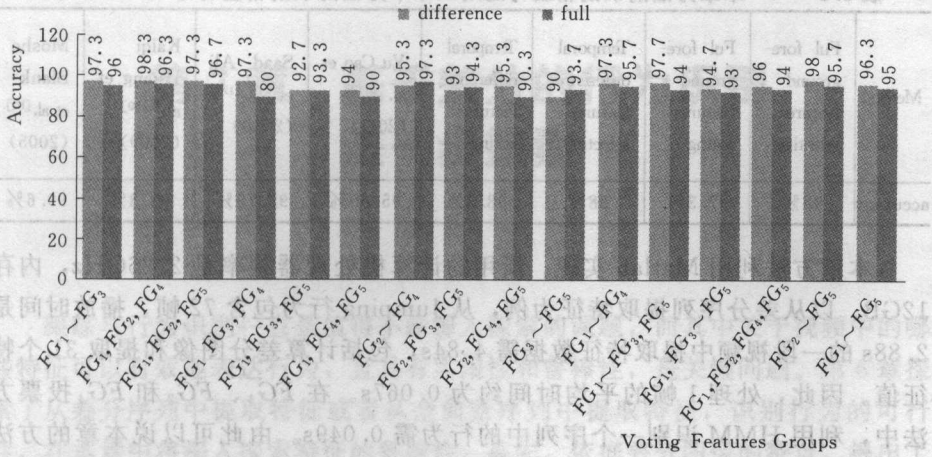


图 6.10 来源于两种序列的特征群投票识别行为结果对比

	Bend	Jack	Jump	Pjump	Run	Side	Skip	Walk	Wave1	Wave2
Bend	100%									
Jack		100%								
Jump			100%							
Pjump				100%						
Run					96.7%		3.3%			
Side						100%				
Skip					6.7%		90%	3.3%		
Walk						3.3%		96.7%		
Wave1									100%	
Wave2										100%

图 6.11 FG_1，FG_2 和 FG_4 投票方法运动 30 次获取的混淆矩阵

将本章的工作效果与近期的一些研究结果进行比较，如表 6.1 所示。表中的测试数据均来源于 Weizmann 行为数据库。可以看到，Blank 等人的工作识别精度很高，他们的方法提取数据时需解 Poisson 等式，这个过程很耗时。从数字上讲，本章方法的识别精度与现时流行方法精度相当，而且只利用了差分序列三方面表观特征即可获取精度 98.3%，这与其他方法比较，有一定优势。

表 6.1　　本章方法的识别精度与近期工作中方法的识别精度对比

Method	Ful fore-ground features selecting	Ful fore-ground features voting	Temporal difference features selecting	Temporal difference features voting	Yu Cao et al. [178] (2011)	Saad Ali et al. [179] (2010)	Kaiqi Huang et al. [180] (2009)	Moshe Blank et al. [171] (2005)
accuracy	98%	97.3%	98%	98.3%	95.60%	95.75%	93.3%	99.6%

本章方法利用 Matlab 实现。所用的计算机处理器频率是 2.26GHz，内存 12GB。以从差分序列提取特征为例，从 Jumping 行为包含 72 帧、播放时间是 2.88s 的一段视频中提取特征数据需 4.84s，包括计算差分图像和提取 31 个特征值。因此，处理 1 帧的平均时间约为 0.067s。在 FG_1、FG_2 和 FG_4 投票方法中，利用 HMM 识别一个序列中的行为需 0.049s。由此可以说本章的方法达到了实时识别。

6.6　讨论与结论

从全前景序列和从差分序列中提取了 10 个方面的表观特征，利用属性选择和投票的方法进行了分析。本章工作的作用有：

(1) 提出了一个基于表观的特征集，并验证了其对行为识别的有效性。

(2) 通过比较静观特征对行为识别的效力，得到一些结论：

1) 具有方向性质的形状特征对行为的表征能力好于单纯的形状特征。

2) 通常联合更多的属性可以提高识别精度，但不是联合的属性越多识别精度越高。所以，精心选择属性是重要的。

(3) 利用属性选择方法和投票方法比较了两类视频源的特征对行为的表征能力。实验结果表明，基于差分序列的特征对行为的表征能力稍强。结论说明以差分序列为源探索行为识别方法的可行性，这是本章工作的意义所在。

当然，Weizmann 数据库是一个纯净的行为数据库，没有缩放、没有光线变化、没有遮蔽和其他噪声，要提高鲁棒性，本章的方法还需改进。另外，基于差分的方法可能对群体行为的识别有优势，因为在群体行为视频中全景个体提取更加困难。

第7章　基于差分的行为识别进一步探索

跟踪与行为识别一直是值得不断深入探索的课题，而其中关于视频中的哪些特征可以有效地表达行为，且具有实用性和鲁棒性，是关键问题。第6章探索了从差分序列中提取特征或者从全前景序列中提取特征，识别行为的可行性，在本章中将深入探索特征的鲁棒性。首先，依据差分图像的特点，提出了计算差分序列中连续两帧光流方法；其次，提出了一个基于表观和光流的特征集；最后，利用HMM进行建模和识别。实验测试在Weizmann行为数据库和KTH行为数据库上进行。

7.1　相关工作介绍和本章方法概述

受一些现实应用的驱动，2D视频中的行为识别研究如火如荼。当然，另一原因是现有研究方法的鲁棒性和实用性还不适应真实场景的应用。人类具有在复杂场景中快速识别其中的行为与事件的能力，这与人类具有学习、记忆、推理的能力相关。让计算机具有像人一样的视觉观察与分析能力一直是计算机视觉研究的长远目标。

第6章工作表明表观特征可以有效表达人的行为，但进一步的测试表明表观特征抗噪性差，在复杂场景中相同方法的识别精度明显下降。本章中将联合表观特征、特征点和光流特征进行行为识别，以提高方法的鲁棒性，进一步证明以差分序列为源做行为识别的可行性。

在相关工作中，表观特征已经在第6章作了介绍，以下介绍兴趣点和光流计算的研究情况。兴趣点可以概括表达视频特征，具有紧凑性的特点，对缩放不敏感，在有光线变化与遮蔽情况下也可以提取。Ivan Laptev等人基于Harris和Forstner兴趣点，识别视频中的事件[145]。在文献［182］中，时空兴趣点作为底层特征，之后建立一种多层统计特征模型来表达行为。

光流技术是另一种用于行为识别中的技术，光流作为一种底层特征用于计算机视觉的方法中。经典的变分方法由Horn和Schunck提出，其方法假设在视频的相邻帧中像素的灰度改变很小，并且进行全局平滑，可以生成密集的光

流[186]。但是变分方法可能令运动的边界变模糊了，也可能在平滑紧密的区域产生空洞。因此人们进行了一些改进，例如：采用各向异性的平滑方法或者采用高阶变分等。由于人是多关节的、复杂动力驱动的生物体，人类的运动可能产生短期大位移的运动，变分方法可能因此失效。Lu 和 Liu 利用 Harris 特征点来弥补变分方法的不足[188]。Brox 等人利用一种基于弯折（warping - based）限定的方法来增强关节运动的表示，同时将复杂的描述符（rich descriptors）整合到变分方法中来捕获大位移运动[190, 198]。另一些研究者利用点路径模型（A point trajectory, PT）来跟踪大位移的运动[194]。PT 模型也可以用来精化半透明场景中光流的偏差[41]。基于块匹配的光流计算方法也可以处理大位移问题[107, 192]。

人类的视觉行为由运动定义，当然运动特征是最终决定行为的因素。从广义上讲，时空特征是静态特征在时间上的运动而产生的。本章中，基于运动的特征主要是指基于光流场计算的特征。光流是人类感觉运动的方式，而人类如何感觉光流从而识别行为，这在计算机领域还没有确切完整的模型[179,180,183]。因为光流是高维的向量场，直接建模与识别有困难，一些研究工作借助如主成分分析（Principal Component Analysis, PCA）等方法来对光流场降维[179, 196]。而欧氏距离不适合表达光流向量间的距离，有的研究工作中，在将 PCA 用于对光流场降维时进行了改进[196]。Ali 等人从光流场序列中提取一些动力学特征，之后利用一种称为 Snapshot PCA 的方法进行降维，来获取运动的动力学模型[179]。

将光流场直接降维获得的主要是场中的能量特征，而人类从光流中感受到的不只是能量。Ali 等人从光流中提取了一系列直观有效的特征来描述行为，特征包括：散度（divergence）、旋度（vorticity）、对称性和不对称性（symmetric and asymmetric）、梯度张量（gradient tensor）、应力和旋转张量（rate of strain and spin tensor）[179]。Ahmad 等联合了运动区域特征和光流特征来描述行为[163]。Huang 等人也联合了光流和其他特征来识别行为[180]。

本章方法的综合过程如图 7.1 所示。行为建模与识别的步骤总结为：

（1）利用相邻帧相减和阈值估计获取帧差序列。每个帧差图像是一个二值图像。

（2）计算连续差分帧间的光流，获取光流场序列。

（3）从差分帧中提取表观特征数据，从光流帧中获取运动特征数据。

（4）利用特征数据训练获取 HMM 模型。

（5）对于一个新的视频段，利用前述步骤中的方法获取其中的特征数据，利用前述获取的 HMM 模型计算新特征数据的似然值，识别新数据所属的行为类别。

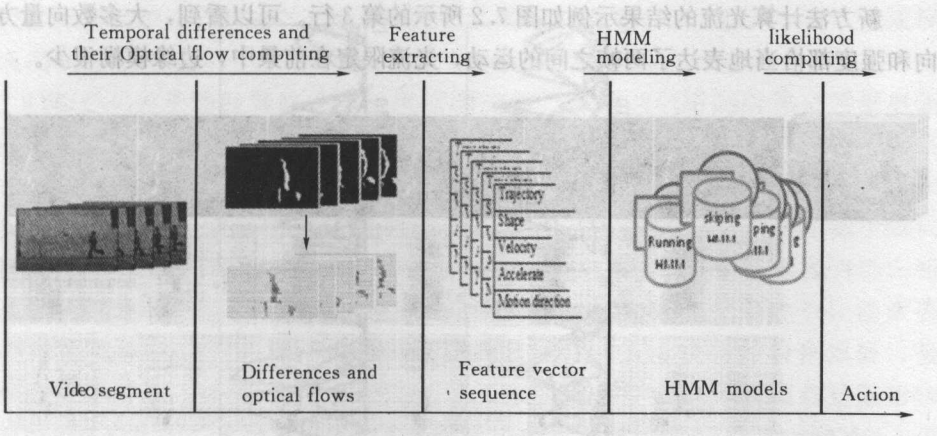


图 7.1 本章方法行为建模与识别过程示意图

本章的方法联合了表观特征和运动特征。与前述工作相比，不同点在于：

(1) 本章方法中特征来源于差分序列。

(2) 联合了差分序列中的表观特征和差分光流序列中的运动特征，与第 6 章的结果进行对比，将验证运动特征相对于表观特征的鲁棒性。

7.2 差分光流计算方法

本章方法基于第 6 章描述的差分图像，前述的光流计算方法用于差分图像会有一些限制，例如：

(1) 差分图像是二值图像，所有的前景灰度和所有的背景灰度均相同，在灰度匹配时产生误差的可能性增加了。

(2) 即使引入各向异性的平滑策略，平滑后边缘模糊现象也很严重。

(3) 差分图像中前景成分很少，利用全局计算方法运算浪费严重。

我们可以从图 7.2 中看到这些限制。因此，本章中利用一个新方法来计算差分序列中连续两帧的光流。新方法描述如下：

(1) 匹配两帧中前景的质心，设全部前景像素与质心运动相同，产生最初的光流场。

(2) 将第 1 帧和第 2 帧图像划分为均匀的块，如 9×9 的块。计算两帧中对应块的前景质心位移，对于质心有位移的块，设整个块中的前景像素均与其质心运动相同；对于质心没有位移的块，保持其中像素的初始光流向量。这个过程中忽略不包含前景像素的块。

(3) 计算两帧中的 Harris 角点[193]，匹配角点，相应调整角点的光流向量。

(4) 在前景范围内平滑光流向量。

新方法计算光流的结果示例如图 7.2 所示的第 3 行。可以看到，大多数向量方向和强度都恰当地表达了两帧之间的运动，光流限定在前景中，边缘模糊很少。

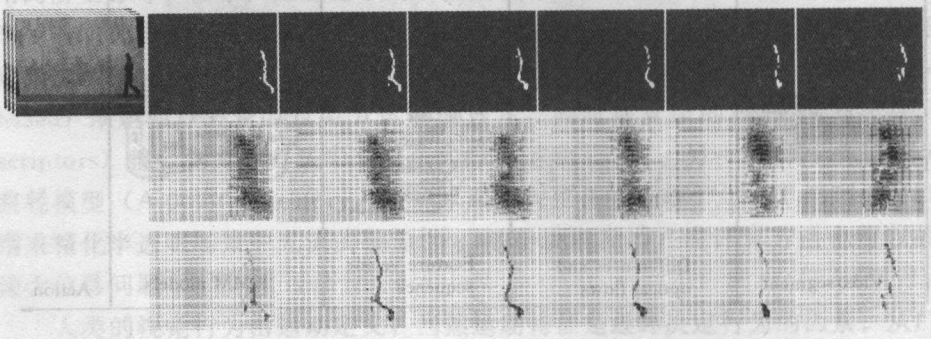


图 7.2　差分图像序列（第 1 行）、利用 Horn－Schunck 方法计算的光流场序列（第 2 行）和用新方法计算的光流场序列（第 3 行）

7.3　特征集

本章利用的特征一部分是从光流序列中提取的运动特征，一部分是从差分序列中提取的表观特征。

参考了人类对光流的直观认识，从光流中提取了若干统计特征，用 $M=(m_1, m_2, \cdots, m_k)$ 来表示。下标 1，2，…，k 只用来标识特征序号，不表示特征的顺序。光流场的向量以（u，v）表示。第 6 章用到的正立的最小外接矩形（UMBR）仍然被用来标准化一些特征值，以减少缩放的影响。

（1）四方向运动速度 。运动方向与速度是光流场中最直观的特征。人类可以感知总体的运动方向。

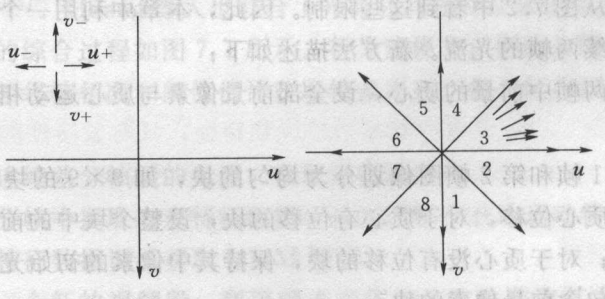


图 7.3　光流的（u，v）坐标和四方向、八方向划分示意图

（a）光流的（u，v）坐标和四方向示意图；（b）光流的（u，v）坐标和八方向示意图

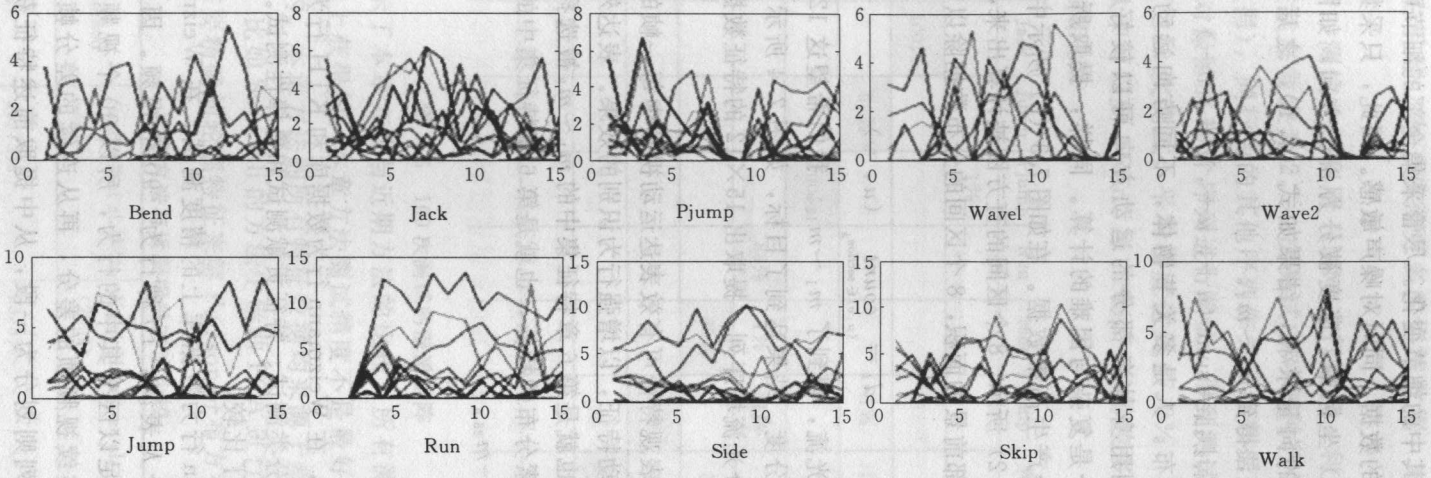


图 7.4　从各行为视频中采集 $m_1 \sim m_{12}$ 散点图

和速度，也可以感知其中的细节变化。尽管采集全部的细节可能精确地表达运动，但也会产生复杂的数据，同时对噪声敏感。因此，只采集了一些方向的速度统计值。在（u，v）坐标中，速度被分为4个方向，如图7.3（a）所示。光流场中4个方向的向量求和，结果如式（7.1），之后，合计值被除以UMBR的对角线长度，得到$m_1 \sim m_4$。

$$v_{1,2,3,4} = \text{sum}\ (u+,\ u-,\ v+,\ v-) \tag{7.1}$$

（2）运动方向分布。人是多关节物体，不同方向运动的分布可以描述行为的一个方面。利用身体各部分的运动方向可以较好地描述运动。但是，提取身体的部分是复杂与困难的计算。同样，提取所有方向的运动分布也是复杂的，且会产生高维数据。在如图7.3（b）所示中，360°方向被划分为8个区间，如式（7.2）所示，8个区间的直方图被计算出来，以粗略表示运动方向分布。设A是全部前景的面积，8个区间的分布数据除以A来进行标准化，得到$m_5 \sim m_{12}$。

$$d_k = \underset{(u,v)\in \text{section}_k}{count}\ (u,\ v) \tag{7.2}$$

至此，基于运动光流，得到了$m_1 \sim m_{12}$。我们希望这12个值的向量对不同的运动有一定的区分度。结果达到了目标，如图7.4所示，图中每个行为，利用了15长度的差分光流场序列，提取出15×12的特征数据，显示在一个散点图中。

（3）表观特征。表观特征可有效表达运动视频中每一帧的姿态，选择了一些表观特征，联合运动特征，以增强行为识别的效果。表达外形伸缩的8方向质心到边界的距离，也就是第6章特征集中的$m_1 \sim m_8$被选择作为本章特征的$m_{13} \sim m_{20}$。8方向像素分布的特征，也就是第6章特征集中的$m_9 \sim m_{16}$被选择作为本章特征的$m_{21} \sim m_{28}$。

7.4　实验与讨论

利用前述的特征，在Weizmann行为数据库和KTH行为数据库上进行了实验测试。实验包括分类测试、特征贡献测试和鲁棒性测试。实验结果也与现有相关的研究结果做了比较。

（1）在Weizmann行为数据库上的精度测试。在Weizmann行为数据库上的每种行为有9个人表演，10种行为，90个视频。因为测试表明长度为15左右的视频段足以区分其中的行为，所以90个视频被划分为更多的视频段。首先计算连续视频帧的差分，再从连续的差分帧计算光流，每生成15长度的光流序列则划分为一段，从中提取前述特征获得28×15的特征数据。人的运动是对称的，例如：从左边走向右边和从右边走向左边，

或者挥左手和挥右手等，为了获得更多的训练数据，将特征左右对称翻转获得双倍的数据。

测试仍遵循留一测试，仍利用 HMM 进行建模。利用高斯混合模型，模型数设为 3，隐藏状态数设为 2。一次测试中对每种行为保留一个特征序列（28×15 的特征数据），同行为的其他序列作为训练数据建立 HMM 模型；获得 10 个模型后，对保留的 10 个序列进行测试，得到测试精度。为了减少随机因素的影响，做了 100 次测试，平均精度达到 97.2%，其中细节如图 7.5 所示的混淆矩阵。从中可以到“Run”与“Skip”混淆较多，同时也可以看到这两个差分序列也很相似。

	Bend	Jack	Jump	Pjump	Run	Side	Skip	Walk	Wave1	Wave2
Bend	99%			1%						
Jack		100%								
Jump			94%				1%	5%		
Pjump				100%						
Run					88%	1%	11%			
Side						97%		3%		
Skip			2%		3%		94%	1%		
Walk								100%		
Wave1	1								100%	
Wave2										100%

图 7.5　100 次测试的混淆矩阵

表 7.1 中显示了本章方法与近期方法的比较，所有测试结果均是在 Weizmann 行为数据库上获得的。本章方法测试精度不是最好的，但与其他方法的结果是有可比性的。本章方法的所有特征数据均来源于差分序列，并且特征数目也较少。另外，比如 Saad Ali 的方法中借助 PCA 进行降维操作，在 Moshe Blank 的方法中需要精确的轮廓特征来解 Poisson 分布方程。

表 7.1　本章方法与近期相关方法比较

	Our method	Saad Ali et al. [179] (2010)	Kaiqi Huang et al. [180] (2009)	Moshe Blank et al. [171] (2005)
classification accuracy	97.2%	95.75%	93.3%	99.6%

续表

	Our method	Saad Ali et al.[179] (2010)	Kaiqi Huang et al.[180] (2009)	Moshe Blank et al.[171] (2005)
Information bases	Temporal - difference images and their optical flow	optical flow	Optical flow combined with foreground frames	Space - time volumes which are computed from foreground silhouettes
features	Velocity, motion direction distribution, shape and pixel distribution	Divergence, vorticity, symmetric andasymmetric fields, gradient tensor, rate of strain and spin tensor	Trajectory, shape, valid pixel portion, average speed, majority direction portion, variance of direction distribution, divergence of direction distribution	Space - time saliency, space - time orientations and weighted moments

本章的测试程序用 Matlab 实现。运行的计算机 CPU 频率 1.73GHz，内存 4GB。处理一段包含 42 帧的“跑”的视频用时 58.14s，过程包括求差分、计算光流和提取特征数据。也就是说处理一帧的时间是 1.38s。另外的测试表明处理“弯腰”视频的一帧需 0.74s。利用获取的 10 个 HMMs，识别一个特征序列的时间是 0.017s。

(2) 在 Weizmann 行为数据库上的特征贡献测试。为了测试不同特征对行为识别的贡献，将特征数据划分开，形成不同成分的子集。利用子集进行识别测试，测试精度结果显示如图 7.6 所示。x 轴显示特征组合，y 轴显示对应识别测试的精度。可以看出，只利用表观特征就可以达到全部特征的识别精度 97.2%。似乎从光流场中提取的运动特征对行为识别精度没有贡献，不过接下来的实验结果中就可以看到运动特征对识别过程的鲁棒性的贡献。

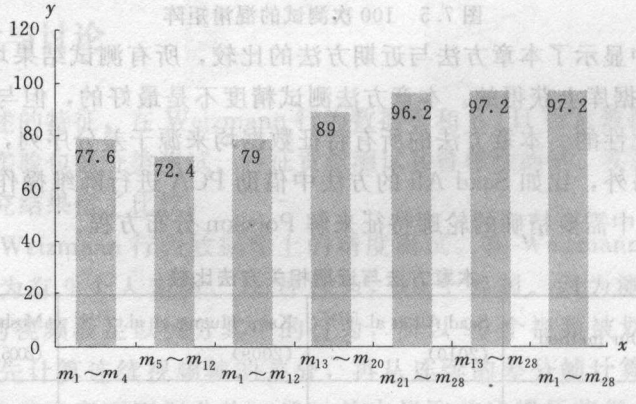


图 7.6 不同特征组合的子集在 Weizmann 行为数据库上测试精度

(3) 在 Weizmann 特殊行为数据库上的鲁棒性测试。在 Weizmann 行为数据库中有一些特殊动作的“走”的视频。这些视频有两类，一类是“走”在水平方向不同视角的视频，角度有 0°, 5°, 10°, 15°, 20°, 25°, 30°, 35°, 40°, 45°；另一类是非常规“走”的视频，例如：走的时候，后面跟一个狗（_dog），走的时候，有一个箱子遮蔽了脚（_nofeet）等。利用前述正常“走”的视频特征训练 HMM 模型，利用这两类特殊“走”视频特征进行测试。测试时，将特征集划分为表观特征（m_{13}～m_{28}）、运动特征（m_1～m_{12}）和全部特征（m_1～m_{28}），测试结果如表 7.2 所示。

测试结果表明，运动特征表达行为的鲁棒性优于表观特征，将两类特征联合起来后，运动特征的鲁棒性一般可以保留下来。

表 7.2　　在特殊“走”行为视频上的精度测试结果

On view change sequences				On irregular sequences			
Test sequences	Test accuracy by using			Test sequences	Test accuracy by using		
	appearance features	motion features	all features		appearance features	motion features	all features
Walking in 0°	50%	100%	100%	_normal	0	0	0
Walking in 5°	50%	75%	75%	_briefcase	50%	100%	100%
Walking in 10°	50%	100%	100%	_nofeet	33.3%	100%	100%
Walking in 15°	60%	80%	80%	_skirt	66.7%	100%	100%
Walking in 20°	40%	80%	40%	_moonwalk	0	100%	75%
Walking in 25°	20%	60%	40%	_limp	83.3%	83.3%	83.3%
Walking in 30°	25%	50%	25%	_dog	0	0	100%
Walking in 35°	25%	0	0	_bag	100%	33.3%	100%
Walking in 40°	0	33%	0	_pole	0	0	0
Walking in 45°	0	20%	0	_kneesup	0	0	0

(4) 在 KTH 行为数据库上的测试。KTH 行为数据库是另一个在行为分析领域中应用较广泛的数据库[146]。第 5 章的实验测试也采用的是这个数据库。其中的视频有 6 种行为，分别是：boxing、handclapping、handwaving、jogging、running 和 walking。每个行为有 25 个人表演，表演的背景有 4 种情况，分别是不同服装、背包、光线变化和视角缩放变化，所以这个数据库是可以测验方法鲁棒性的视频。利用本章的方法进行行为识别测试，结果如图 7.7 所示。

如表 7.3 所示显示了本章方法与近期其他方法识别精度的对比。可以看出，本章方法的识别精度与其他方法是相当的。当然，如果在方法中增加精细的去噪过程会提高识别精度，因为观察视频直接计算得到的差分图像，可以看到明显的噪声。

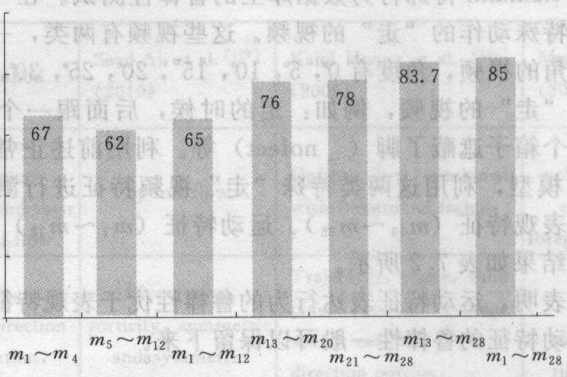


图 7.7　在 KTH 数据库上的测试结果

表 7.3　　在 KTH 视频数据库上的识别精度对比

Method	C. chuldt et al. [37] (2004)	Saad Ali et al. [1] (2010)	Our method
Accuracy	71.72%	87.7%	85%

当然，Weizmann 和 KTH 均是流行的行分析数据库，一些方法在其上的测试已经达到了很高的精度[179,180,195,196]，文中表 7.2 与表 7.3 中选择的对比方法与本章方法是有一定可比性的。例如：Rahman 等人的方法在 KTH 上的测试精度达到了 94.67%，但是其测试只选择了 KTH 视频中“情景 1”的子集，这个子集是全部视频中噪声最小的[195]，因此与我们的方法的结果不具有可比性。

7.5　结论

本章提出了一种完全基于差分序列的行为识别方法。在这些方法中联合使用了基于运动的特征和基于表观的特征。方法的识别精度与近期一些研究工作的识别精度相当。本章工作的贡献有：

(1) 基于视频差分序列，提出了一个联合运动与表观的特征集。利用这个特征集，行为识别精度可以达到与近期一些研究工作的识别精度相当。这个结果表明了基于视频差分序列进行行为识别的可行性。

(2) 提出了一种适应差分图像特征的光流计算方法。

(3) 通过不同特征子集在两个视频数据库上的贡献测试，说明基于运动的特征表达行为的鲁棒性，联合表观特征可以增强整体特征的鲁棒性且不降低表观特征的表征能力。

当然，本章的方法是基于全局特征的，全局特征对遮蔽很敏感。人类可以利用局部的少量特征识别行为，进一步深化人类的感知特点，开发更鲁棒和实用的方法是计算机视觉持久努力的方向。

第8章 结 论

数字技术的广泛应用和数据存储技术的快速进步形成了海量数据积累。数据挖掘是一类数据分析技术，它将传统的数据分析方法与处理大量数据的复杂算法相结合。其任务包括频繁项集挖掘、关联规则挖掘、聚类、分类、特异数据挖掘和时间序列挖掘等。随着网络与多媒体技术的发展，数据的形式更加多样化，数量日益增大，这对数据挖掘算法的研究和数据挖掘与领域知识技术的融合都提出了新的挑战。本书的研究工作以数据挖掘算法分析与改进为基础，分析了相关算法在智能视频监控领域的应用构架，对行为识别的视频特征进行了分析与验证，具体内容有：

(1) 研究了利用聚类来简化全局特异数据挖掘计算的方法。特异数据被定义为在数据集中只被少数对象拥有，并且与其他数据显著不同的数据。特异数据挖掘算法有两类：一类是基于密度的局部特异数据挖掘；一类是基于距离的全局特异数据挖掘。Zhong Ning 等人提出了一种基于距离的全局特异数据挖掘算法，计算一个数据的特异程度，要计算其与其他所有数据的距离之和，其计算时间复杂度为 $O(N^2)$，其中，N 是数据集的势。如果先将数据集根据距离聚类，聚类后数据较多的类和距离数据集平均值很近的类，其类中的每一个数据都不可能成为特异数据，将这样的类按其均值整体参加运算。只有很少的小类中的数据每一个单独参加运算，这样可将运算复杂度降低到 $O(n^2)$，其中，n 是单独参加运算的数据个数与整体参加运算的大类的类个数之和。由特异数据的特点，n 可能远小于 N，从而大大提高了运算效率。实验表明，基于聚类的特异数据挖掘算法其挖掘特异数据能力强于 Zhong Ning 教授的原算法，与基于密度的局部特异数据挖掘算法的能力相当，时间效率显著好于 Zhong Ning 教授的算法。

(2) 研究了规则“$C \rightarrow A$”在分类中的作用。分类是有监督的机器学习与识别过程，C4.5 算法、CBA 算法等被归类为基于规则的分类算法。这类算法的共同点是采集和利用形如“$A \rightarrow C$”的规则特征，其中，A 表示全部或部分条件属性的一些取值组成的集合，C 表示某个类标号。在逻辑上，“$A \rightarrow C$”表示 A 对 C 的支持，而“$C \rightarrow A$”表示 A 对 C 的必要，在分类中如果能利用“$C \rightarrow A$”的特征也许会提高分类精度。按这样的想法，编制了两种实验方法，方法 1 只考虑“$A \rightarrow C$”的影响，方法 2 考虑“$A \rightarrow C$”和“$C \rightarrow A$”的影响。

两种方法时间复杂度均为线性的。分别在 UCI 机器学习库的 5 个分类集 Mushroom、Wine、Zoo、Breast 和 KDDCUP99 网络访问记录数据集上进行了实验测试。结果表明，如果能采集到合适的“$C \rightarrow A$”规则，并让其在分类中起作用，可以有效提高分类精度。并且，对不平衡数据集，只有小类的规则被采集和利用，具有生成规则集小，训练与测试时间线性的优势。

(3) 分析了视频监控技术现状，提出一种智能监控系统的构架，提出并验证一类行为识别的视频特征。监控系统构架中数据由原始录像的视频文件、模式和实时数据 3 层构成。从原始录像中提取数据形成特征模式，由特征识别行为，由行为确定异常。观察人运动的 2D 视频，不同的运动行为在一定程度上表现为人身体不同部位的伸与缩。将人运动前景矩形在横和纵方向上划分为均匀的区间，采集这些区间的宽度及其内部空档变化的序列，以序列的频率和时间平均方差构成特征向量。为了验证此特征对行为识别的有效性，采用线性判别式方法、支持向量机方法、k 最近邻方法、线性参数分类方法等模式识别方法，进行了分类交叉检验、特征值分析，进行了不同粗细划分的特征数据识别精度对比，进行了不同视频分段的识别精度对比。并对数据集进行了线性判别分析与特异分析。实验结果表明，当视频分段长度达到一定值，区间划分达到一定精细程度时，利用特征数据能有效识别不同的行为，特征向量的各分量在分类中均有效。并且，特征数据线性可分性较好，具有较好的“不同类别间的特征值距离较远，同一类别内的特征距离较近”特性。

(4) 从全前景序列和从差分序列中提取了 10 个方面的表观特征，利用属性选择和投票的方法进行了分析。证明了以下结论：

1) 具有方向性质的形状特征对行为的表征能力好于单纯的形状特征。

2) 基于差分序列的特征对行为的表征能力稍强。

结论说明以差分序列为源探索行为识别方法的可行性，这是这部分工作的意义所在。进一步联合差分序列的运动特征提高了表达行为的鲁棒性，且不降低表观特征的表征能力。

当然，计算机视觉技术还在发展，适应复杂场景的行为分析方法还有待深入探索。本书的方法是基于全局特征的，全局特征对遮蔽很敏感。另外，基于差分的方法可能对群体行为的识别有优势，因为在群体行为视频中全景个体提取更加困难。人类可以利用局部的少量特征识别行为，进一步深化人类的感知特点，开发更鲁棒和实用的方法是计算机视觉持久努力的方向。

参考文献

[1] 企博网. 国外部分有影响力的数据挖掘软件列表 [EB/OL]. (2005-10-5) [2006-12-5]. http://www.bokee.net/forummodule/view/ForumThread/ view/10/74217.html.

[2] 中国数据库联盟论坛. 数据挖掘软件产品 [EB/OL]. (2006-9-18) [2006-11-5]. http://www.dmgroup.org.cn.

[3] Margaret H Dunham. 数据挖掘教程 [M]. 郭崇慧, 译. 北京: 清华大学出版社, 2005.

[4] Kosala R., Blockeel H.. Web mining research: A survey [C]. In: SIGKDD Explorations - Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining, 2000, 2.

[5] 韩家炜, 孟小峰, 王静, 等. Web 挖掘研究 [J]. 计算机研究与发展, 2001, 38 (4): 405-414.

[6] Piotr S. Szczepaniak. Web Intelligence Research - Activity within the Polish Center [C]. In: Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence, 2004.

[7] Frank Zhigang Wang, Sheng Jiang, Yau Jim Yip. Web Intelligence Research Activities in the UK [C]. In: Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence, 2004.

[8] 王本年, 高阳, 陈世福, 等. Web 智能研究现状与发展趋势 [J]. 计算机研究与发展, 2005, 42 (5): 721-727.

[9] 段其国, 苗夺谦, 陈敏, 等. 计算 Web 智能研究综述 [J]. 计算机科学, 2007, 34 (7): 1-4.

[10] 雅虎. IMatch 搜索引擎上线贴近用户实际需求 [EB/OL]. (2006-7-19) [2008-1-9]. http://cn.tech.yahoo.com/ 060719/639/2hwkc.html.

[11] Malik Agyemang, Ken Barker, Reda Alhajj. Framework for mining web content outliers [C]. In: Proceedings of the 2004 ACM symposium on Applied computing, 2004, 3.

[12] Malik Agyemang, Ken Barker, Reda Alhajj. WCOND-mine: algorithm for detecting Web content outliers from Web documents [C]. In: Proceedings of 10th IEEE Symposium on Computers and Communications, 2005, 6.

[13] Tao Jiang, Ah-Hwee Tan, Ke Wang. Mining Generalized Associations of Semantic Relations from Textual Web Content [J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 19 (2): 164-179.

[14] Ramakrishna Varadarajan, Vagelis Hristidis, and Tao Li. Beyond Single-Page Web Search Results [J]. IEEE Transactions on Knowledge and Data Engineering, 2008, 20 (3): 411-424.

[15] Wingyan Chung, Hsinchun Chen, Nunamaker J. F. Jr. Business intelligence explorer: a knowledge map framework for discovering business intelligence on the Web [C].

In: Proceedings of the 36th Annual Hawaii International Conference on System Sciences, 2003.

[16] Xin Chen, Yi fang Brook Wu. Web mining from competitors' websites [C]. In: Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, 2005.

[17] Lee R. S. T., Liu J. N. K. iJADE Web-miner: an intelligent agent framework for Internet shopping [J]. IEEE Transactions on Knowledge and Data Engineering, 2004, 16 (4): 461-473.

[18] Yufei Li, Yuan Wang, Xiaotao Huang. A Relation-Based Search Engine in Semantic Web [J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 19 (2): 273-282.

[19] Jang Minsu, Sohn Joo-Chan, Cho Hyun Kyu. Automated Question Answering Using Semantic Web Services [C]. In: the 2nd IEEE Asia-Pacific Service Computing Conference, 2007, 12: 344 - 348.

[20] Mukherjea S., Bamba B. and Kankar, P. Information retrieval and knowledge discovery utilizing a biomedical patent semantic Web [J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17 (8): 1099-1110.

[21] 人民网. 关注企业第四核心竞争力企业情报不是007 [EB/OL]. (2006-4-8) [2008-2-5]. http://www.people.com.cn/GB/jingji/1045/22483 44.html.

[22] 邓三鸿，杨建林，潘有能，等. 企业门户网站中的数据挖掘研究 [J]. 情报学报，2003，22 (1)：40-45.

[23] 杨建林，孙明军. 竞争情报收集的自动化 [J]. 情报杂志，2005，1：40-43.

[24] 百度. 百度eCIS5.0 [EB/OL]. (2007-3-8) [2008-2-12]. http://hi.baidu.com/fomy/blog/item/150621fa476d0c899e5146c6.html.

[25] Jiawei Han, Micheline Kamber. Data Mining: Concepts and Techniques [M]. Second Edition. Beijing: China Machine Press, 2006.

[26] 焦李成，刘芳. 智能数据挖掘与知识发现 [M]. 西安：西安电子科技大学出版社，2006.

[27] 毛国君，段立娟. 数据挖掘原理与算法 [M]. 北京：清华大学出版社，2005.

[28] R Agrawal, T Imielinski, A Swami. Mining association rules between sets of items in large databases [C]. ACM SIGMOD Int' l Conf. Management of Data, Washington D. C., 1993.

[29] Jiawei Han, Jian Pei, Y Yin. Mining frequent patterns without candidate generation [C]. ACM SIGMOD Int' l Conf. Management of Data, Dallas, TX, 2000, 5: 1-12.

[30] Y. Chi, Y. yang and R. R. Muntz. Indexing and mining free trees [C]. IEEE Int' l Conf. on Data Mining, 2003.

[31] Y. Chi, Y. yang and R. R. Muntz. Mining frequent rooted trees and free trees using canonical forms: technical report [R]. CSD-TR No. 030043, 2004.

[32] Y. Chi, Y. yang. CMTreeMiner: Mining both closed and maximal frequent subtrees: technical report [R]. CSD-TR No. 030053, 2004.

[33] J. S. Park. An effective hash - based algorithm for mining association rules [C]. SIGMOD 1995.

[34] R. Agarwal, C. Aggarwal, and V. V. V. Prassad. A Tree Projection Agorithm for Generation of Frequent Itemsets [J]. Journal of Parallel and Distributed Computing, 2000.

[35] 宋余庆，朱玉全. 基于 FP - tree 的最大频繁项目集挖掘及更新算法 [J]. 软件学报，2003，14 (9)：1586 - 1592.

[36] 陈慧萍，王建东，叶飞跃. MAXFP - Miner：利用 FP - tree 快速挖掘最大频繁项集 [J]. 控制与决策，2005，20 (8)：887 - 891.

[37] 吉根林，赵斌，孙志挥. 利用 Hash 树生成频繁项目集的新方法 [J]. 小型微型计算机系统，2004，25 (10)：1841 - 1843.

[38] J. Pei, J. Han, and R. Mao. CLOSET：An efficient algorithm for mining frequent closed itemsets [C]. ACM SIGMOD Int' l Workshop on Data Mining and Knowledge Discovery, Dallas, TX, 2000, 5：11 - 20.

[39] M. J. Zaki and C. J. Hsiao. CHARM：An efficient algorithm for closed association rule mining：technical report [R]. Rensselaer Polytechnic Institute, October 1999.

[40] 陈耿，朱玉全，宋余庆，等. 基于频繁模式树的约束最大频繁项目集挖掘算法研究 [J]. 应用科学学报，2006，24 (1)：64 - 69.

[41] 宋余庆，朱玉全，杨鹤标. 一种基于频繁模式树的约束最大频繁项目集挖掘及其更新算法[J]. 计算机研究与发展，2005，42 (5)：777 - 783.

[42] D. I. Lin and Z. M. kedem. Pincer - search：A new algorithm for discovering the maximum frequent set [C]. The 6th Intl. Conf. Extending Database Technology, March 1998.

[43] Burdick D, Calimlim M, Gehrke J. Mafia：A maximal frequent itemset algorithm for transactional databases [C]. In：Proc. of the 17th Int' l Conf. on Data Engineering, 2001.

[44] Gouda K, Zaki MJ. Efficiently mining maximal frequent itemsets [C]. In：Proc. of the 1st IEEE Int' l Conf. on Data Mining, 2001.

[45] Wang H, Li QH. An improved maximal frequent itemset algorithm [C]. In：Proc. the 9th Int' l Conf. on the Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, 2003.

[46] Grahne G, Zhu JF. High performance mining of maximal frequent itemsets [C]. In：Proc. of the 6th SIAM Int' l Workshop on High Performance Data Mining, 2003：135 - 143.

[47] Li Yang. Pruning and Visualizing Generalized Association Rules in Parallel Coordinates [J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17 (1).

[48] M. J. Zaki, Chingjui Hsiao. Efficient Algorithms for Mining Closed Itemsets and Their Lattice Structure [J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17 (4).

[49] Jianyong Wang, Jiawei Han, Ying Lu, et al. TFP：An Efficient Algorithm for Mining Top - K Frequent Closed Itemsets [J]. IEEE Transactions on Knowledge and Data

Engineering, 2005, 17 (5) .

[50] Murat Kantarcioglu and Chris Clifton. Privacy - Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data [J] . IEEE Transactions on Knowledge and Data Engineering, 2005, 16 (9) .

[51] Vassilios S. Verykios. Association Rule Hiding [J] . IEEE Transactions on Knowledge and Data Engineering, 2005, 16 (4) .

[52] Gosta Grahne, Jianfei Zhu. Fast Algorithms for Frequent Itemset Mining Using FP - Trees [J] . IEEE Transactions on Knowledge and Data Engineering, 2005, 17 (10) .

[53] Zhou QH, Wesley C, Lu BJ. SmartMiner: A depthfirst algorithm guided by tail information for mining maximal frequent itemsets [C] . In: Proc. of the IEEE Int' l Conf. on Data Mining, 2002: 570 - 577.

[54] M. J. Zaki, K. Gouda. Fast vertical mining using diffsets: Technical report [R] . Rensselaer Polytechnic Institute, March 2001.

[55] Jiuyong Li, Hong Shen, Rodney Topor. Mining the smallest association rule set for predictions [C] . In the proceedings of the 2001 IEEE international conference on data mining, 2001: 361 - 368.

[56] 吉根林，韦素云．分布式环境下约束性关联规则的快速更新 [J] ．东南大学学报（自然科学版），2006，36 (1)：35 - 38.

[57] R. Ng, L. V. S. Lakshmanan, J. Han, et al. Exploratory mining and pruning optimizations of constrained association rules [C] . In Proc. ACM SIGMOD Int. Conf. Management of Data, Seattle, WA, 1998, 6: 13 - 24.

[58] 杨文杰，胡明昊，唐振民，等．一种有效的基于约束的关联规则发现算法 [J] ．南京理工大学学报，2005，29 (1)：109 - 113.

[59] 董雁适，程翼宇，潘云鹤．基于高频模式树的项约束关联规则发现方法 [J] ．浙江大学学报（工学版），2002，36 (7)：445 - 450.

[60] 寇育敬，王春花，黄厚宽．约束关联规则的增量式维护算法 [J] ．计算机研究与发展，2001，38 (8)：947 - 951.

[61] 卢炎生，杨芬，赵栋．带单调约束的关联规则挖掘 [J] ．计算机工程，2004，30 (8)：78，79，129.

[62] 高飞，谢维信．发现含有第一类项目约束的频繁集的快速算法 [J] ．计算机研究与发展，2001，38 (11)：1295 - 1301.

[63] 崔立新，苑森淼，赵春喜．约束性相联规则发现方法及算法 [J] ．计算机学报，2000，23 (2)：216 - 220.

[64] R. J. Bayardo, R. Agrawal and D. Gunopulos. Constraint based rule mining on large, dense data sets [C] . In Proc. 1999 Int' l. Conf. Data Engineering, Sydney, Australia, Apr. 1999.

[65] S. Brin, R. Motwani and C. Silverstein. Beyond market basket: Generalizing associa-

tion rules to correlations [C]. In Proc. 1997 ACM SIGMOD Int' l. Conf. Management of Data, Tucson, Arizona, 1997, 5: 265 - 276.

[66] G. Grahne, L. Lakshmanan and X. Wang. Efficient mining of constrained correlated sets [C]. In Proc. 2000 Int. Conf. Data Engineering, San Diego, CA, 2000, 2: 512 - 521.

[67] L. V. S. Lakshmanan R. Ng, J. Han and A. Pang. Optimization of constrained frequent set queries with 2 - variable constraints [C]. In Proc. 1999 ACM SIGMOD Int. Conf. Management of Data, Philadelphia, PA, 1999, 6: 157 - 168.

[68] J. Pei and J. Han. Can we push more constraints into frequent pattern mining? [C]. In Proc. of Int. Conf. Knowledge Discovery and Data Mining, Boston, MA, 2000, 8: 350 - 354.

[69] R. Srikant, Q. Vu, and R. Agrawal. Mining association rules with item constraints [C]. In Proc. of Int. Conf. Knowledge Discovery and Data Mining, Newport Beach, CA, 1997, 8: 67 - 73.

[70] Raymond T. Ng, Laks V. S. Lakshmenan, Jiawei Han, et al. Exploratory Mining and Optimization of Constrained Association Queries: Technical Report [R]. University of British Columbia and Concordia University, October 1997.

[71] Carlos Ordonez and Edward Omiecinski. Efficient Disk - Based K - Means Clustering for Relational Databases [J]. IEEE Transactions on Knowledge and Data Engineering, 2004, 16 (8): 909 - 921.

[72] Carlos Ordonez. Integrating K - Means Clustering with a Relational DBMS Using SQL [J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18 (2): 188 - 201.

[73] S. Sarawagi, S. Thomas and R. Agrawal. Integrating Mining with Relational Databases: Alternatives and Implications [C]. In: Proc. of ACM SIGMOD Conf. 1998.

[74] T. Imielinski and A. Virmani. MSQL: A Query Language for Database Mining [J]. Data Mining and Knowledge Discovery, 1999, 3 (4): 373 - 408.

[75] H. Jamil. Ad Hoc Association Rule Mining as SQL3 Queries [C]. In: Proc. of IEEE int' l Conf. Data Mining, 2001: 609 - 612.

[76] R. Meo, G. Psaila and S. Ceri. An Extension to SQL for Mining Association Rules [J]. Data Mining and Knowledge Discovery, 1998, 2 (2): 195 - 224.

[77] S. Chaudhuri. Data mining and database systems: Where is the intersection? [J]. Data Engineering Bulletin, 1998, 21: 4 - 8.

[78] T. Imielinski and H. Mannila. A database perspective on knowledge discovery [J]. In: proc. of CACM 1996.

[79] A. Silberschatz and S. Zdonik. Database systems - breaking out of the box [J]. SIGMOD Record, 1997, 26: 36 - 50.

[80] 张卫丰，丁艺明．结构化挖掘语言 SML 的设计与实现 [J]．小型微型计算机系统，2001，22 (2)：195 - 198.

[81] 杨炳儒，孙海洪，熊范纶．利用标准 SQL 查询挖掘多值型关联规则及其评价[J]．计算机研究与发展，2003，39（3）：307－312.

[82] Han J，Fu Y，Koperski K. DMQL：A data mining query language for relational databases [J]．Data Mining and Knowledge Discovery，1996，1：461－465.

[83] Han J，Fu Y，Wang W. DBMiner：a system for mining Knowledge in large relational databases [J]．Data Mining and Knowledge Discovery，1996，20：250－255.

[84] Y. Fu and J. Han. Meta－rule－guided mining of association rules in relational databases [C]．In proc. of first Int' l Workshop on Intergration of Knowledge Discovery with Deductive and OO－Databases，Singapore，1995：39－46.

[85] Rosa Meo，Giuseppe Psaila and Stefano Ceri. A new SQL－like operator for mining association rules [C]．In proceedings of the 22nd VLDB Conference，Mumbai，India，1996.

[86] Haixun Wang，Carlo Zaniolo，Chang Richard Luo. ATLAS：a small but complete SQL extension for data mining and data streams [C]．In proceedings of the 29th VLDB Conference，Berlin，Germany，2003.

[87] Imielinski T，Virmani A and Abdulghani A. Discovery board application programming interface and query language for database mining [C]．In Proc. of the 2nd Int' l Conference on Knowledge Dicovery and Data Mining，Portland，Oregon，August 1996.

[88] Lavington S，Dewhurst N，Wilkins E and Freitas A. Interfacing knowledge discovery algorithms to large database management systems [J]．Information and Software Technology，1999：605－617.

[89] Jian Pei. Pattern－growth methods for frequent pattern mining（phd thesis）[D]．Directed by Jiawei Han. Canada：SIMON FRASER UNIVERSITY，2002.

[90] Sergios Theodoridis，Konstantinos Koutroumbas. 模式识别 [M]．李晶皎，王爱侠，张广渊，等，译．北京：电子工业出版社，2006.

[91] 李弼程，邵美珍，黄洁，等．模式识别原理与应用 [M]．西安：西安电子科技大学出版社，2008.

[92] 孙即祥．现代模式识别 [M]．北京：高等教育出版社，2002.

[93] 边肇祺，张学工．模式识别（2 版）[M]．北京：清华大学出版社，2000.

[94] 张云涛，龚玲．数据挖掘原理与技术 [M]．北京：电子工业出版社，2004.

[95] 朱玉全，杨鹤标，孙蕾．数据挖掘技术 [M]．武汉：东南大学出版社，2006.

[96] 陈尚勤，等，模式识别 [M]．成都：成都电讯工程学院出版社，1985.

[97] Pang－Ning Tan，Michael Steinbach and Vipin Kumar. 数据挖掘导论 [M]．范明，范宏建，译．北京：人民邮电出版社，2006. 7.

[98] Quinlan J. R. C4. 5：Programs for Machine Learning [R]．Los Altos：Morgan Kaufmann，1993.

[99] Liu B.，Hsu W. and Ma Y. Integrating Classification and Association Rule Mining [C]．In the Proceeding of the Fourth International Conference on Knowledge Discovery and Data Mining，New York，1998：80－86.

[100] P. Domingos，M. Pazzanil. Beyond independence：Conditions for the optimality of the simple Bayesian classifier [C]．In the proceeding of the 13th International Con-

ference on Machine Learning, San Francisco, 1996.

[101] Muneaki Ohshima, Ning Zhong, Yiyu Yao, et al. Relational peculiarity - oriented mining [J]. Data Mining and Knowledge Discovery, 2007, 15: 249 - 273.

[102] Zhong N, Yao YY, Ohshima M, et al. Interestingness, peculiarity, and multi - database mining [C]. In: Proceedings of the 2001 IEEE international conference on data mining, 2001: 566 - 573.

[103] Zhong N, Yao YY, Ohshima M. Peculiarity oriented multidatabase mining [J]. IEEE Transactions on Knowledge and Data Engineering, 2003, 4: 952 - 960.

[104] Zhong N, Liu C, Yao YY, et al. Relational peculiarity oriented data mining [C]. In: Proceedings of the 2004 IEEE international conference on data mining, 2004: 575 - 578.

[105] D. Hawkins. Identification of outliers [R]. Chapman and Hall, London, 1980.

[106] M. M. Breuning, H. P. Kriegel, R. T. Ng, et al. Lof: Identifying density _ based local outliers [C]. In proc. 2000 ACM SIGMOD Int. Conf. Management of Data, Dallas, Texas, 2000.

[107] He Zengyou, Xu Xiaofei, Deng Shengchun. Discovering Cluster Based Local Outliers [J]. Pattern Recognition Letters, 2003, 24 (9 - 10): 1651 - 1660.

[108] E. Knorr and R. Ng. Algorithms for mining distance - based outliers in large datasets [C]. In Proc. 1998 Int. Conf. Very Large Data Bases, New York, Aug. 1998: 392 - 403.

[109] E. Knorr and R. Ng. A unified notion of outliers: properties and computation [C]. In Proc. KDD 1997: 219 - 222.

[110] E. Knorr, R. Ng, V. Tucakov. Distance - based outliers: algorithms and applications [J]. The VLDB Journal, 2000, 8 (3/4): 237 - 253.

[111] He Z, Xu X, Deng S. Squeezer: An efficient algorithm for clustering categorical data [J]. Journal of Computer Science and Technology, 2002, 17 (5): 611 - 625.

[112] 李庆华，李新，蒋盛益．一种面向高维混全属性数据的异常挖掘算法 [J]．计算机应用，2005，25 (6)：1353 - 1356.

[113] Zhou Aoying, Wei Li, Yu Fang. Effective discovery of exception class association rules [J]. Journal of computer science and technology, 2002, 5: 304 - 313.

[114] W Li, J Han, J Pei. CMAR: Accurate and efficient classification based on multiple class - association rules [C]. In proc. 2001 Int. Conf. Data Mining, San Jose, CA, 2001: 369 - 376.

[115] X Yin, J Han. CPAR: Classification based on predictive association rules [C]. In Proc. of international conference on data mining, San Francisco, 2003, 5: 331 - 335.

[116] Blake CL and Merz CJ. UCI Machine Learning repository of machine learning databases [DB/OL]. (1987 - 1 - 5) [2006 - 3 - 4]. http: //archive. ics. uci. edu/ml/.

[117] Hettich S. and Bay S. D. The UCI KDD Archive [DB/OL]. (1987 - 1 - 5) [2006 - 2 - 10]. http: //kdd. ics. uci. edu.

[118] 洪流．国外安防技术新进展 [J]．中国安防产品信息，2000，6.

[119] 张浩，蔡晋辉，黄平捷，等．基于贝叶斯统计推理的复杂场景边缘检测 [J]．华南

理工大学学报（自然科学版），2007，35（9）：40－44.

[120] 罗敏，朱晓岷，李小红，等．基于径向小波变换的图像特征提取算法［J］．武汉大学学报（信息科学版），2008，33（1）：29－32.

[121] 代科学，李国辉，涂丹，等．监控视频运动目标检测减背景技术的研究现状和展望［J］．中国图像图形学报，2006，11（7）：919－927.

[122] Mun Wai Lee，Isaac Cohen. A Model－Based Approach for Estimating Human 3D Poses in Static Images［J］. IEEE Transactions on Pattern Analysis and Machine Intelligence，2006，28（6）：905－916.

[123] J. K. Aggarwal，Sangho Park. Human Motion：Modeling and Recognition of Actions and Interactions［C］. In proceedings of the Second International Symposium on 3D Data Processing，Visualization and Transmission，2004：640－647.

[124] Saeid Rahati，Reihaneh Moravejian，Ehsan Mohamad Kazemi，et al. Vehicle Recognition Using Contour let Transform and SVM［C］. In the proceeding of Fifth International Conference on Information Technology：New Generations，2008：894－898.

[125] 张继平，刘直芳．视频中运动目标的实时检测和跟踪［J］．计算机测量与控制，2004，12（11）：1036－1051.

[126] Ismail Haritaoglu，David Harwood，Larry S. Davis. W4：Real－Time Surveillance of People and Their Activities［J］. IEEE Transactions on Pattern Analysis and Machine Intelligence，2000，22（8）：809－830.

[127] Oberli C，Urzua J，et al. An expert system for monitor alarm integration［J］. J Clin Monit 1999，15：29－35.

[128] PC PRO. AI could power next－gen CCTV cameras［EB/OL］.（2008－3－2）［2008－6－25］. http：//www. pcpro. co. uk/news208452ai－could－power－nextgen－cctv－cameras. html.

[129] Sony. Intelligent IP Video Monitoring Based on the Sony Unique Distributed Enhanced Processing Architecture（DEPA）Platform Incorporates Intelligent Video Analysis to Provide a High－Level of Security at an Affordable Price［EB/OL］.（2007－1－15）［2007－5－12］. http：//www. sonybiz. net/ nvm.

[130] 中国科学院深圳先进技术研究院［EB/OL］.［2007－2－15］. http：//www. siat. ac. cn/project2. php.

[131] 中国科学院计算技术研究所．白春礼出席高清动态智能监控系统应用签约仪式［EB/OL］.（2007－6－27）［2007－7－10］. http：//www. cas. ac. cn/ html/ Dir/ 2007/06/27/ 15/09/83. html.

[132] 北京智安邦科技有限公司［EB/OL］.［2007－7－1］. http：//www. zanb. com. cn/ newEbiz1/ EbizPortalFG/ portal/html/ index. html.

[133] Panasonic. Security Products［EB/OL］.［2007－5－12］. http：//www. panasonic. com/business/ security.

[134] Vidient［EB/OL］.［2007－5－1］. http：//www. vidient. com/.

[135] VistaScape Security System［EB/OL］.［2007－2－19］. http：//www. vistascape. com/.

[136] ACM portal [EB/OL]. [2007-4-2]. http://portal.acm.org/citation.cfm?id=1178782.

[137] PETS: Performance Evaluation of Tracking and Surveillance [EB/OL]. (2005-11-4) [2007-2-2]. http://www.cvg.cs.rdg.ac.uk/slides/pets.html.

[138] 代科学，付畅俭，武德峰，等. 视频挖掘：概念、技术与应用 [J]. 计算机应用研究，2006，1：1-4.

[139] 胡芝兰，江帆，王贵锦，等. 基于运动方向的异常行为检测 [J]. 自动化学报，2008，34 (11)：1348-1357.

[140] 徐光祐，曹媛媛. 动作识别与行为理解综述 [J]. 中国图象图形学报，2009，14 (2)：189-195.

[141] Catherine Achard, Xingtai Qu, Arash Mokhber, et al. A novel approach for recognition of human actions with semi-global features [J]. Machine Vision and Applications, 2008, 19: 27-34.

[142] Lena Gorelick, Moshe Blank, Eli Shechtman, et al. Actions as Space-Time Shapes [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29 (12): 2247-2253.

[143] Bobick AF, Davis JW. The recognition of human movement using temporal templates [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, 23 (3): 257-267.

[144] Ivan Laptev and Tony Lindeberg. Local Descriptors for Spatio-Temporal Recognition [C]. In: Proceedings of ECCV Workshop on Spatial Coherence for Visual Motion Analysis, 2004: 1-12.

[145] Ivan Laptev and Tony Lindeberg. Velocity adaptation of space-time interest points [C]. In: Proceedings of ICPR'04, Cambridge, UK, 2004: 52-56.

[146] Christian Schuldt, Ivan Laptev and Barbara Caputo. Recognizing Human Actions: A Local SVM Approach [C]. In: Proceeding of ICPR'04, Cambridge, UK, 2004: 32-36.

[147] Chung P C, Liu C D. A daily behavior enabled hidden Markov model for human behavior understanding [J]. Pattern Recognition, 2008, 41 (5): 1572-1580.

[148] 贾克斌，邓智毗，庄新月. 基于时序特征的视频相似性匹配算法 [J]. 北京工业大学学报，2008，34 (12)：1250-1253.

[149] Lj L Buturovic. PCP: a program for supervised classification of gene expression profiles [J]. Bioinformatics, 2005, 11: 245-247.

[150] C. Chang and C. Lin. LIBSVM: a library for support vector machines [R]. 2001.

[151] 余建英，何旭宏. 数据统计分析与 SPSS 应用 [M]. 北京：人民邮电出版社，2003.

[152] Rudra N. Hota, Vijendran Venkoparao and Anupama Rajagopal. Shape based Object Classification for Automated Video Surveillance with Feature Selection [C]. in Proc. 10th International Conference on Information Technology, 2007, pp. 97-99.

[153] A. F. Bobick and J. W. Davis. The Recognition of Human Movement Using Temporal Templates [J]. IEEE Trans. on PAMI, Vol. 23, No. 3, pp. 257-

267, 2001.

[154] Ronald Poppe. A survey on vision - based human action recognition [J] . Image and Vision Computing, 2010, 28 (6): 976 - 990.

[155] Xin Xu, J. Tang, Xiao ming, Liu, et al. Human behavior understanding for video surveillance: Recent advance [C] . In Proc. 2010 IEEE Intl. Conf. on Systems Man and Cybernetics (SMC), pp. 3867 - 3873.

[156] Joshua Candamo, Matthew Shreve, Dmitry B. Goldgof, Deborah B. Sapper, and Rangachar Kasturi. Understanding transit scenes: a survey on human behavior - recognition algorithms [J] . IEEE Trans. Intelligent Transportation Systems, Vol. 11, pp. 206 - 224, Mar. 2010.

[157] J. K. Aggarwal and M. S. Ryoo. Human activity analysis: a review [J] . ACM Computing Surveys, April 2011, Vol. 43, No. 3, Article 16, pp. 1 - 43.

[158] Caroline Rougier, Jean Meunier, Alain St - Arnaud, et al. Fall Detection from Human Shape and Motion History using Video Surveillance [C] . In 21st international networking and applications workshops, May, 2007, pp. 875 - 880.

[159] Hiroki Murayama and Keiichi Yamada. Detection of Unusual Human Activity based on Sequence of Actions with MHI and CDP [C] . In 2010 IEEE region 10 conference, Nov. 2010, pp. 1663 - 1667.

[160] Ju Han and Bir Bhanu. Individual Recognition using Gait Energy Image [J] . IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 28, No. 2, pp. 316 - 322, 2006.

[161] Chen Wang, Junping Zhang, Liang Wang, et al. Human Identification Using Temporal Information Preserving Gait Template [C] . IEEE Transactions on Pattern Analysis and Machine Intelligence, has been accepted for publication, 2011.

[162] Joveria Javed, Hashim Yasin and Syed Faisal Ali. Human Movement Recognition using Euclidean Distance: A Tricky Approach [C] . In the 3rd International Congress on Image and Signal Processing , CISP2010, pp. 317 - 321.

[163] Mohiuddin Ahmad and Seong - Whan Lee. HMM - based human action recognition using multi - view image sequences [C] . In proc. 18th International Conference on Pattern Recognition, 2006, pp. 263 - 266.

[164] N. Robertson andI. Reid. A general method for human activity recognition in video [C] . Computer Vision and Image Understanding, pp. 232 - 248, 2006.

[165] C. Stauffer and W. E. L. Grimson. Adaptive background mixture models for real - time tracking [C] . In Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog. , 1999, vol. 2, pp. 246 - 252.

[166] Michael Harville. A framework for high - level feedback to adaptive, per - pixel, Mixture - of - Gaussian background models [C] . In the 7th European Conference on Computer Vision (ECCV - 2002), May, 2002, Copenhagen, Denmark. Lecture Notes in Computer Science, 2002, Volume 2352/2002, pp. 37 - 49.

[167] Dahjye Lee, Pengcheng Zhan, Aaron Thomas, et al. Shape - based human intrusion detection [C] . SPIE International Symposium on Defense and Security, Visual In-

formation Processing XIII, Vol. 5438, pp. 81 - 91, Orlando, Florida, USA, April 12 - 16, 2004.

[168] Hao Jiang, Mark S. Drew, and Zenian Li. Successive convex matching for action detection [C]. In Proc. IEEE CS Conf. Computer Vision and Pattern Recognition, 2006.

[169] Fu yuan Hu, Yan ning Zhang, Lan Yao. An effective detection algorithm for moving object with complex background [C]. In Proc. IEEE Intl. Conf. on Machine Learning and Cybernetics, 2005, Vol. 8, pp. 5011 - 5015.

[170] Yan Yuqin, Song Shijun, He Shujuan, et al. Feature matching algorithm of moving human bodies [C]. In Proc. Intl. Conf. on Information Technology and Computer Science, 2009, Vol. 2, pp. 217 - 220.

[171] Moshe Blank, Lena Gorelick, Eli Shechtman, et al. Actions as space - time shapes [C]. In Proc. Tenth IEEE International Conference on Computer Vision (ICCV), 2005.

[172] L. Gupta, M. Srinath. Invariant planar shape recognition using dynamic alignment [C]. In Proc. IEEE Int' l Conf. on Acoustics, Speech, and Signal Processing, Vol. 12, pp. 217 - 220, 1987.

[173] Weiming Hu, Min Hu, Xue Zhou, et al. Principal axis - based correspondence between multiple cameras for people tracking [J]. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 28, No. 4, pp. 663 - 671, APRIL 2006.

[174] C. Stauffer and E. Grimson. Learning patterns of activity using real time tracking [J]. IEEE Trans. Pattern Anal. Mach. Intell., Vol. 22, No. 8, pp. 747 - 757, Aug. 2000.

[175] Minghui Hu. Visual pattern recognition by moment invariants [J]. IRE Trans. On Information theory, Vol. 8, pp. 179 - 187, 1962.

[176] Matthew Brand and Vera Kettnaker. Discovery and segmentation of activities in video [J]. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 22, No. 8, pp. 844 - 851, Aug. 2000.

[177] Kevin Murphy. Bayes net toolbox for Matlab [OL]. Availabel: http: // www. cs. ubc. ca/~murphyk /Software/HMM/hmm _ usage. html.

[178] Yu Cao, Lili Ju, Qin Zou, et al. A unified framework for locating and recognizing human actions [C]. In CVPR, 2011.

[179] Saad Ali and Mubarak Shah. "Human action recognition in videos using kinematic features and multiple instance learning," IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 32, pp. 288 - 303, Feb. 2010.

[180] Kaiqi Huang, Shiquan Wang, Tieniu Tan, et al. Maybank. Human behavior analysis based on a new motion descriptor [J]. IEEE Trans. On Circuits and Systems for Video Technology, Vol. 19, pp. 1830 - 1840, Dec. 2009.

[181] Ivan Laptev, Tony Lindeberg. Space - time interest points [C]. In: Proc. of Ninth IEEE International Conference on Computer Vision, 2003, pp. 432 - 439.

[182] Jun Yin, Yan Meng. Human activity recognition in video using a hierarchical probabi-

listic latent model [C] . In: proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2010, pp. 15 - 20.

[183] Seyed Ali Etemad, Pierre Payeur, Ali Arya. Automatic temporal location and classification of human actions based on optical features [C] . In: proc. 2nd International Congress on Image and Signal Processing, 2009, pp. 1 - 5.

[184] Yuichi Motai, Sumit Kumar Jha, Daniel Kruse. Human tracking from a mobile agent: Optical flow and Kalman filter arbitration [J] . Signal Processing: Image Communication 27 (2012) 83 - 95.

[185] Aaron Bobick, James Davis. An appearance - based representation of action [C] . In: Proc. IEEE CS Conf. Computer Vision and Pattern Recognition, 1996, pp. 307 - 312.

[186] Berthold Horn, Brian Schunck. Detemining optical flow [J] . Artificial Intelligence 17 (1981) 185 - 204.

[187] B. D. Lucas, T. Kanade. An iterative image registration technique with an application to stereo vision [C] . In: Proc. Imaging Understanding Workshop, 1981, pp. 121 - 130.

[188] Lu Ziyun, Liu Wei. The compensated HS optical flow estimation based on matching Harris corner points [C] . In: Proc. International Conference on Electrical and Control Engineering (ICECE), 2010, pp. 2279 - 2282.

[189] Nils Papenberg, Andres Bruhn, Thomas Brox, et al. Highly accurate optical flow computation with theoretically justified warping [C] . Int' l J. Computer Vision 67 (2006) 141 - 158.

[190] Thomas Brox, Jitendra Malik. Displacement optical flow: descriptor matching in variational motion estimation [J] . IEEE Trans. On PAMI 33 (2011) 500 - 513.

[191] Michael J. Black, P. Anandan. The robust estimation of multiple motions: parametric and piecewise smooth flow fields [J] . Computer Vision and Image Understanding 63 (1996) 75 - 104.

[192] Bernd Kitt, Benjamin Ranft, Henning Lategahn. Block - matching based optical flow estimation with reduced search space based on geometric constraints [C] . In: Proc. 13th International IEEE Conference on Intelligent Transportation Systems (ITSC), 2010, pp. 1104 - 1109.

[193] Chris Harris, Mike Stephens. A combined corner and edge detector [C] . In: Alvey Vision Conference, 1988, pp. 147 - 151.

[194] P. Sand, S. Teller. Particle video: Long - range motion estimation using point trajectories [C] . Int. J. Comput. Vis. 80 (2008) 72 - 91.

[195] S. A. Rahman, S. Y. Cho, M. K. H. Leung. Recognising human actions by analyzing negative spaces [C] . IET Computer Vision 6 (2012) 197 - 213.

[196] Shoushun Chen, Polina Akselrod, Bo Zhao, et al. Efficient Feedforward Categorization of Objects and Human Postures with Address - Event Image Sensors [J] . IEEE Trans on PAMI 34 (2012) 302 - 314.

[197] Hidetomo Sakaino. A Semitransparency - Based Optical - Flow Method With a Point

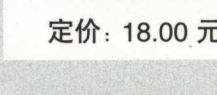
Trajectory Model for Particle - Like Video [J] . IEEE Trans on Image Processing 21 (2012) 441 - 450.

[198] T. Brox, A. Bruhn, N. Papenberg, et al. High accuracy optical flow estimation based on a theory for warping [C] . In: Proc. ECCV, 2004, pp. 25 - 36.

[199] Sirinart Tangruamsub, Keisuke Takada, Osamu Hasegawa. A fast online incremental learning method for object detection and pose classification using voting and combined appearance modeling [J] . Signal Processing: Image Communication, 27 (2012) 75 - 82.

责任编辑：武丽丽 魏素洁 帅丹

销售分类：算法



定价：18.00 元